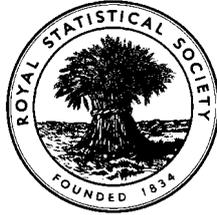# EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 1997

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. Annual data relating to the United Kingdom, 1973-1994, have been entered into a Minitab worksheet as follows:

C1 contains the Year 1973, 1974, ..., 1994

C2 contains Gross National Product, £m at 1990 prices

C7 contains Gross Domestic Fixed Capital Formation, £m at 1990 prices.

(Source: United Kingdom National Accounts, 1995 edn, table 1.3)

They have been analysed as shown on the computer output on the page overleaf of this examination paper.

[Note that the effect of "MTB > delete entry 1 in C5" is to raise the entries in C5 by one space, so that for example the number which would otherwise have been in row 3 is in row 2.]

(i)   What do C5 and C10 contain? Discuss the statistical technique which they exemplify.

(ii) Use C6 and C11 to draw a time chart with a logarithmic vertical axis showing both Gross National Product and Gross Domestic Fixed Capital Formation. Why do statisticians sometimes draw such time charts with logarithmic vertical axes?

(iii) Use the computer output and your time chart to write a substantial economic analysis of Gross National Product and Gross Domestic Fixed Capital Formation in the United Kingdom over the period 1973 to 1994.

Turn over

A2. A random sample of fifteen firms listed in the Building and Construction section of the Financial Times share prices section on 24 May 1996 was drawn, and their earnings per share (pence) as given in the 1994-1995 Stock Exchange Yearbook for the two most recent years were noted as follows:

	Previous year	*Most recent year*
Abbey	4.1	14.5
AMFC	-44.5	3.6
Berkeley	16.0	33.6
Bryant	4.9	8.9
CRP Leisure	-6.6	-1.9
Clarke (T)	-0.2	7.3
Costain	-84.6	18.8
Crest Nicholson	-4.0	0.0
Gleeson (MJ)	67.1	57.5
Hewden-Stuart	3.7	6.8
Laing, J	9.4	15.0
Taylor Woodrow	-22.4	4.1
Tilbury Douglas	-2.0	46.2
Vibroplant	2.7	3.8
Wimpey (George)	-40.1	6.8

Denoting the previous year figures as x and the most recent year figures as y, it may be found that $\sum x = -96.5$, $\sum x^2 = 16219.35$, $\sum y = 225.0$, $\sum y^2 = 7631.14$ and $\sum xy = 2478.81$.

Draw a scatter diagram showing the above data.

Find and test for statistical significance the ordinary (product-moment) correlation coefficient and the rank correlation coefficient between the earnings per share in the two years.

Use the t distribution to test the null hypothesis that, in the population of all building and construction firms, the means of earnings per share in the two years were the same.

Without doing any further calculations, explain one non-parametric test which might be used to examine possible changes in the profitability of firms in this industry. Would it have any advantages over the test which you carried out?

A similar analysis of data from a random sample of fifteen investment trust companies gave an ordinary correlation coefficient of 0.979 and a rank correlation coefficient of 0.950. The null hypothesis that the means in the two years were the same was accepted ($p = 0.83$). Discuss the different features of the two types of company as revealed by this analysis.

Turn over

A3.

**Logarithms of consumers' purchases of jewellery etc
and total consumers' purchases, United Kingdom, 1984-1994,**

	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
y	4.21	4.22	4.27	4.36	4.48	4.60	4.61	4.61	4.52	4.47	4.39
x	4.34	4.38	4.44	4.49	4.57	4.60	4.61	4.58	4.58	4.61	4.64

(Source: derived from United Kingdom National Accounts, 1995 edn, tables 4.7, 4.8)

Note: The above data are natural logarithms of Laspeyres quantity indexes, in each case with $1990 = 100$. y refers to jewellery etc and x to total consumers' purchases.

From the above data it may be found that $\Sigma y = 48.74$, $\Sigma x = 49.84$, $\Sigma y^2 = 216.1930$, $\Sigma x^2 = 225.9250$ and $\Sigma xy = 220.9676$.

The equation $y_t = \alpha + \beta x_t + u_t$, where the u's are Normally distributed homoscedastic and serially independent disturbance terms, has been proposed as a model for the data. Prove that, in this model, β is the elasticity of purchases of jewellery etc with respect to total expenditure.

Fit the model, and estimate standard errors of the coefficients, together with $r^2(xy)$, the coefficient of determination. Find a 95 per cent confidence interval for the elasticity.

Writing $t = -5$, -4 … 5 for the years shown, it may be found that $r(xt) = 0.90172$ and $r(yt) = 0.65139$.   Hence find and test for statistical significance the partial correlation coefficient $r(xy.t)$ and the coefficient of multiple determination $R^2(y.xt)$.

What do these coefficients tell about the multiple regression of y on x and t?

A4. Quarterly statistics of consumers' expenditure on food (y) and total consumers' expenditure (x) at 1990 prices, seasonally adjusted, in millions of pounds, for the United Kingdom for the period 1981 to 1994 are collected from Economic Trends Annual Supplement, 1996 edn. The variable t is a time trend taking values 0, 0.25, ..., 13.75, and $Q_i (i = 2, 3, 4)$ equals 1 in ith quarters and 0 otherwise. Regressions are obtained as follows, with standard errors in parentheses.

$$y = 7103 + 0.03623\,x + 33.30\,t \qquad R^2 = 0.939, \qquad s = 128.5 \qquad\qquad ....(i)$$
$$ (356) \quad (0.00580) \quad (14.29) \qquad \Sigma e_i^2 = 874,685$$

$$y = 7136 + 0.03611\,x + 33.72\,t - 38.01\,Q_2 - 41.63\,Q_3 - 25.15\,Q_4, \quad R^2 = 0.940, \quad s = 131.1$$
$$ (366) \quad (0.00593) \quad (14.60) \qquad (49.59) \quad (49.62) \qquad (49.69) \quad \Sigma e_i^2 = 859,801 \ ...(ii)$$

(a) Why was no Q_1 variable included in (ii)?

(b) Test the statistical significance of the three Q variables as a set.

(c) Explain what one learns about consumers' expenditure on food from these regressions.

(d) Regressions of y on x and t for 1981 to 1986 and for 1989 to 1994 are calculated separately, with the following results:

1981 to 1986
$$y = 4406 + 0.08345\,x - 126.55\,t \qquad R^2 = 0.695, s = 104.2 \quad ....(iii)$$
$$ (1184) \quad (0.001976) \quad (47.56) \qquad \Sigma e_i^2 = 228,189$$

1989 to 1994
$$y = 6631 + 0.03965\,x + 50.40\,t \qquad R^2 = 0.723, \quad s = 89.9 \quad ....(iv)$$
$$ (1014) \quad (0.01243) \quad (12.38) \qquad \Sigma e_i^2 = 169,733$$

Test the null hypothesis that the variance of the stochastic term u_t in the model

$$y_t = \alpha + \beta x_t + \gamma t + u_t \quad \left(t = 0,\ 0.25,...,13.75\right) \qquad\qquad ....(v)$$

was constant over the whole period against an alternative hypothesis to be stated.

(e) What is the economic interpretation of the coefficient γ in the model (v)?

Test the null hypothesis that γ was the same in both of the sub-periods examined in (iii) and (iv) against the two-sided alternative hypothesis that it was not. (You may assume that the numbers of observations were sufficient for you to use large sample methods.)

Turn over

SECTION B - ECONOMETRICS

B1. An economist has estimated by least squares the model

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + \varepsilon_t \quad (t = 1,\ldots,n)$$

from annual data on consumption (y_t) and income (x_t).

(i) Explain how the parameters of this model are interpreted.

(ii) It is suspected that the error terms ε_t might be autocorrelated, through

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$$

where

$$E(\eta_t) = 0, \qquad E(\eta_t^2) = \sigma^2, \quad E(\eta_t\eta_s) = 0 \text{ if } t \neq s.$$

Derive the Lagrange multiplier test of the null hypothesis $\rho = 0$. Also, briefly describe an alternative test of this null hypothesis and outline the principles underlying it.

(iii) A colleague has suggested the alternative loglinear model

$$\log y_t = \alpha + \beta\log x_t + \gamma\log y_{t-1} + \varepsilon_t \quad (t = 1,\ldots,n).$$

Outline the case for the loglinear specification. How would you choose from these two alternative specifications?

B2. In analysing economic time series, analysts generally ask whether individual series are trend stationary (ie $E(x_t) = \mu$ and $\text{cov}(x_t, x_{t+k}) = \gamma(k)$ for all t) or stationary only after first differencing. Discuss fully why it is thought important to ask this question, the approaches that are used to answer the question, and any limitations of these approaches.

B3. Let $\left(Y_{1t}, Y_{2t}\right)$ be a pair of time series, and consider the two-equation model

$$Y_{1t} - Y_{1,t-1} = \varepsilon_{1t},$$

$$Y_{2t} - Y_{2,t-1} = \gamma\left(Y_{1,t-1} - \beta Y_{2,t-1}\right) + \varepsilon_{2t}.$$

Let $\varepsilon_t' = \left(\varepsilon_{1t}, \varepsilon_{2t}\right)$ and assume that

$$E\left(\underset{\sim}{\varepsilon_t}\right) = \underset{\sim}{0} \qquad E\left(\underset{\sim}{\varepsilon_t}\,\underset{\sim}{\varepsilon_t'}\right) = \Omega \qquad E\left(\underset{\sim}{\varepsilon_t}\,\underset{\sim}{\varepsilon_s'}\right) = 0 \qquad \text{if } t \neq s.$$

(i) Find the single series ARIMA models for

    (a) Y_{1t},   (b) $Y_{1t} - \beta Y_{2t}$,   (c) Y_{2t}.

(ii) Discuss as fully as possible the implications of your answers in (i).

B4. Individual i will choose to make a large purchase if the marginal benefit from that purchase is greater than the marginal cost. Let y_i^* denote the difference between marginal benefit and cost, assumed to depend on economic variables $x_{ji}\left(j = 1,..., K\right)$ through

$$y_i^* = \alpha + \beta_1 x_{1i} + ... + \beta_K x_{Ki} + \varepsilon_i,$$

where ε_i is a random error term. As a practical matter, the econometrician cannot observe y_i^*. All that can be observed is whether a purchase was made; that is, whether y_i^* is positive.

(i) Show that, if the distribution of ε_i is continuous and symmetric about zero,

$$P\left(y_i^* > 0\right) = F\left(\alpha + \sum_{j=1}^{K} \beta_j x_{ji}\right),$$

    where F is the cumulative distribution function of ε_i.

(ii) Fully discuss how this model can be analysed, given observable data.

(iii) Outline some other practical applied economic problems for which this model is relevant.

Turn over

SECTION C - OPERATIONAL RESEARCH

C1. (a) Given the three random uniform numbers 0.614, 0.352, 0.887, obtain three random observations from

   (i) the exponential distribution with mean 6,
   (ii) the uniform distribution between 25 and 75.

   (b) What are the antithetic variates corresponding to these observations? For what purpose are antithetic variables used in simulation, and what is the rationale behind their use?

(c) Given the following pairs of uniformly distributed U(0,1) random numbers

(0.096, 0.610)	(0.370, 0.855)
(0.665, 0.506)	(0.912, 0.391)
(0.142, 0.188)	(0.226, 0.206)
(0.484, 0.763)	

use the rejection method to obtain random observations from the distribution with p.d.f. *f(x)*, where

$$f(x) = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 1, \\ \frac{1}{2}(1 - \frac{1}{3}(x-1)), & 1 \leq x \leq 4, \\ 0, & \text{otherwise}, \end{cases}$$

C2. A company can manufacture five products, each of which requires processing on the same three machines. Data are available concerning the processing times (in hours) of each product on each machine, the unit profit on each product, and the total weekly capacity (in hours) of each machine.

The problem of determining the quantity of each product to manufacture in order to maximize the total profit, subject to the capacities of the machines, has been formulated as a linear programme and solved using the simplex method.

The final tableau is given below. The variables x_i $(i = 1, ..., 5)$ represent the quantity of product i that is produced, and s_1, s_2, and s_3 are the slack variables corresponding to the three machine capacity constraints

Basis	x_1	x_2	x_3	x_4	x_5	s_1	s_2	s_3	RHS
x_2	0	1	-1.5	1.625	0.375	0.125	0	-0.075	7.2
s_2	0	0	3	-6.75	-9.25	0.25	1	-0.65	14.4
x_1	1	0	2.5	-0.625	-0.625	-0.125	0	0.125	12.0
z	0	0	125	231.25	368.5	6.25	0	23.75	10920

(i)   What is the optimal production plan for the company, i.e. which products should they make, and how much of each?

(ii)   Is there any spare machine capacity, and if so, for which machine or machines?

(iii)   How much should the company be willing to pay per hour for extra machine time, for any machine or machines for which there is no spare capacity?

(iv)   By how much would the unit profit on Product 3 have to increase in order for that product to be worth making?

(v)   How could you tell whether there are any alternative solutions to this problem? How would you find them?

(vi)   Denoting by p_i the unit profit on product i, $(i = 1, ..., 5)$, by t_{ij} the processing time for product i on machine j, $(i = 1, ..., 5, \ j = 1, ..., 3)$, and by Cap_j the total weekly capacity of machine j, formulate this problem as a linear programme using the variables x_i $(i = 1, ..., 5)$.

(vii)   In the above problem, the values of the machine capacities are $Cap_1 = 288$, $Cap_2 = 192$ and $Cap_3 = 384$. Within what ranges could each of these values change so as to leave the optimum basis unaltered?

Turn over

C3. A project consists of ten activities whose durations are uncertain. The following precedence table shows the estimated mean and variance of each activity.

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

(i) Draw a network diagram for this project.

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

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

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

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

C4. (a) A sports shop obtains two types of basketball from the same supplier. For each type, the price of an individual ball, the cost of placing an order, the holding cost per ball per annum and the annual demand are given in the following table. Demand for both types is steady and shortages must not occur.

Basketball type	Price (£)	Order cost (£)	Holding cost (£)	Demand
1	4	10	5	900
2	15	19	8	304

The supplier gives a discount of 21 % whenever a buyer spends at least £1200 on a single type of ball. Determine the optimal order quantity and the total annual cost for each type of basketball.

(b) Describe the costs involved in maintaining inventory. What is meant by the term *lead time*?

(c) What is the meaning of the values s and S in an (s, S) inventory policy? Describe briefly the rules of this policy. What are the major drawbacks of this approach?

(d) A school operates a snack shop at which pupils may buy chocolate bars at break time. The lead time is one week. The demand in this period is variable, but can be modelled by a normal distribution with mean = 350 and standard deviation 10. The school aims to be able to meet demand in 95% of all cycles. How much safety stock should the school maintain? What is the re-order point?

SECTION D - MEDICAL STATISTICS

D1. The Weibull survival distribution is characterised by the hazard function $\lambda(t) = \lambda\gamma t^{\gamma-1}$

(i) Explain what is meant by the *hazard function*. Define and derive the corresponding survival function, $S(t)$, and probability density function, $f(t)$, for the Weibull distribution. Sketch the form of these three functions for the cases $\gamma = 0.5, 1, 1.5$.

(ii) Evaluate the cumulative hazard function. Describe a graphical method, based upon the estimated cumulative hazard function, for checking whether data appear to come from a Weibull distribution.

(iii) The follow-up/survival times for n randomly selected patients are $t_1, t_2, ..., t_n$. Corresponding indicator variables $I_1, I_2, ..., I_n$ show whether these times are censored ($I_i = 1$) or not ($I_i = 0$). Derive the likelihood function for λ and γ.

Turn over

D2. (a) Define the *sensitivity, specificity, positive predictive value* and *negative predictive value* of a diagnostic test.

(b) According to the W.H.O. definition, patients are considered to have diabetes if their blood glucose concentration is more than 11.1 mmol/l two hours after being given a 75g glucose drink. A study examined the extent to which the blood glucose concentration prior to taking the drink may be used to provide a simpler test for diabetes. 408 men went through the glucose tolerance test so that they could be classified as diabetic or not by the W.H.O. definition.

Initial glucose concentration c(mmol/l)	Not diabetic	Diabetic	Total
$0 < c \bullet 5.5$	101	2	103
$5.5 < c \bullet 6.0$	139	3	142
$6.0 < c \bullet 6.5$	95	4	99
$6.5 < c \bullet 7.0$	29	5	34
$7.0 < c \bullet 8.0$	14	3	17
$8.0 < c$	1	12	13
Total	379	29	408

Five potential cutoff values for a diagnostic test are 5.5, 6.0, 6.5, 7.0 and 8.0 mmol/l. Determine the corresponding test sensitivities and specificities. Determine the positive and negative predictive values of the tests in populations in which the prevalence of the disease is (i) 1% and (ii) 20%.

(c) Sketch the ROC curve using these cutoff values.

D3. W.H.O. regularly publish the volume "Cancer Incidence in Five Continents". This provides age- and sex-specific numbers of cancers and corresponding incidence rates for over thirty sites of cancer from registries all over the world.

(i) Discuss how comparisons of rates could be made between different countries, mentioning difficulties that would need to be addressed. What might be learned from such comparisons?

(ii) Discuss how comparisons of rates could be made within the same cancer registry, but over different periods of time. Again, mention the difficulties that would need to be addressed and the potential benefits of such an exercise.

(iii) Describe the statistical model that is most often used to analyse such tables of counts across age, sex, time and registry.

D4. (a) In a randomised clinical trial n patients received a new treatment and another n received the standard treatment. The probability of survival to one year using the standard treatment is believed to be 90%. The new treatment would be worth introducing regularly if its one-year percentage survival were 94% or higher (i.e. it increased the percentage surviving by at least 4%). Derive an approximate formula for the necessary sample size n in terms of the type I error (α) and type II error (β), using a two-sided test. Evaluate n for $\alpha = 0.05$ and $\beta = 0.2$.

(b) Discuss the advantages and disadvantages of conducting such a trial in several centres.

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

Turn over

E1. An experiment is to be carried out to discover whether covering ornamental plants during the early stages of their growth can allow a reduction in the amount of a nutrient sprayed on to them. Four doses of the spray, with the plants covered and uncovered, (making eight treatments in total) are to be compared. The experiment will be conducted in a glasshouse compartment, which contains six benches, parallel to each other and each running east-west. On each bench two trays of plants can be accommodated, each tray containing four pots of three plants each. It is convenient to cover a tray of plants, rather than individual pots or plants. Different levels of the spray can be applied to different pots within the same tray, but not to different plants within the same pot. There are expected to be large differences in growth measurements between the north and south parts of the glasshouse, but not between east and west.

Suggest a suitable design for this experiment, i.e. state how to determine which level of spray should be applied to each pot, whether each tray should be covered or uncovered and where each tray will be placed. You do not need to randomise the design, but you should explain fully how this should be done.

Outline the analysis of variance which will be calculated for the data collected from this experiment, indicating the numbers of degrees of freedom for each row of the table.

E2. A clinical trial was performed to find the best combination of two anti-hypertensive drugs. All combinations of three equally spaced doses of Drug A and four equally spaced doses of Drug B were tested, with 25 patients in each group, using a completely randomised design. The mean reductions in blood pressure (mm Hg) for the twelve treatments were as follows.

		Drug B			
		0	1	2	3
Drug A	0	2.7	6.4	7.3	5.6
	1	8.1	11.6	10.0	11.0
	2	7.6	8.5	9.0	9.7

In accordance with the trial protocol the data were analysed by fitting a second-order polynomial model. This gave the following results.

Analysis of Variance:

Source of Variation	df	SS	MS
A	1	20.48	20.48
B	1	9.20	9.20
A*A	1	25.22	25.22
B*B	1	5.47	5.47
A*B	1	0.20	0.20
Lack of fit	6	6.42	1.07
Pure error	288	83.13	0.28
Total	299	150.12	

Parameter	Estimate	SE
Intercept	3.44	0.87
A	7.96	1.41
B	2.95	0.99
A*A	-3.08	0.63
B*B	-0.68	0.30
A*B	-0.14	0.33

(i) Interpret these results fully. This should include a recommendation as to the best combination of doses for reducing blood pressure, a sketch of any plots which would be helpful in conveying the results to the clinicians and a brief description of any further analysis which should be carried out.

(ii) Instead of fitting the second-order polynomial, a standard factorial analysis could have been carried out, giving two degrees of freedom for the effect of Drug A, three for the effect of Drug B and six for the interaction. Discuss the advantages and disadvantages of this alternative analysis, without carrying out any detailed calculations.

Turn over

E3. In enzyme kinetic studies, the relationship between enzyme activity, v, and substrate concentration, s, is modelled using the Michaelis-Menten equation, $v = \dfrac{v_{max}\, s}{K + s}$, where v_{max} and K are unknown parameters which have to be estimated. The following data were collected from an experiment.

s	v	s	v
5	1.0	175	13.0
10	1.6	200	11.0
25	2.0	225	13.6
50	4.0	250	13.0
75	6.0	275	15.0
100	10.0	300	15.0
125	9.6	400	18.0
150	10.6	500	20.0

(i) Draw a scatterplot of v against s and by considering the behaviour of v for large values of s, or otherwise, obtain initial estimates of the parameters. Explain how they can be used in fitting the Michaelis-Menten equation by nonlinear least squares. What assumptions would have to be made about the error structure?

(ii) Biochemists often fit the Michaelis-Menten equation by writing it as $\dfrac{1}{v} = \dfrac{K}{v_{max}}\dfrac{1}{s} + \dfrac{1}{v_{max}}$ and performing a simple linear regression of $\dfrac{1}{v}$ on $\dfrac{1}{s}$. Discuss the relative merits of the two procedures.

E4. A bioassay was performed to test the toxicity of a new compound which has the potential to be used in the treatment of an acute condition. Five doses were given to 40 rats each and the numbers of deaths were as follows.

Dose	Number of deaths out of 40
1	5
2	19
4	31
8	34
16	39

It was decided to model the data using $\log\left(\dfrac{p}{1-p}\right) = \beta_0 + \beta_1 \log(d)$, where p is the proportion of rats killed and d is the dose of the compound. Check whether this model seems appropriate by sketching an appropriate plot.

Fitting this model gave the following results:

	Parameter	Estimate	s.e.
1	Intercept	-1.568	0.318
2	log(dose)	1.833	0.255

Scaled deviance = 2.7760 on 3 d.f.

Variance matrix of parameter estimates

1	0.1011	
2	-0.0653	0.0650
	1	2

Estimate the ED50 of the compound and obtain a 95% confidence interval for the ED50 using Fieller's theorem.

SECTION F - STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A company manufactures vehicle exhaust systems, and offers a two-year warranty to replace faulty items. A note is kept of claims each month. They are classified into four categories: perforation of silencer, failure of baffles within silencer, failure of rubber mounting bracket, and other. Results for the last 20 months are given in the table below.

(i) Summarise these data, in the form of a diagram (or diagrams), for presentation to the production manager.

Month	Reason for Claim				Total
	Perforation	*Baffles*	*Bracket*	*Other*	
1	1	4	0	1	6
2	5	2	0	0	7
3	2	1	0	0	3
4	0	6	1	2	9
5	1	2	0	0	3
6	3	3	0	1	7
7	3	5	0	1	9
8	2	5	0	0	7
9	0	3	0	1	4
10	1	4	7	1	13
11	2	1	2	0	5
12	1	3	1	2	7
13	1	7	0	0	8
14	0	2	0	0	2
15	4	2	0	0	6
16	2	4	1	1	8
17	0	6	0	0	6
18	1	3	0	0	4
19	3	8	0	0	11
20	2	5	0	0	7
Totals	34	76	12	10	132

(ii) What other information should the manager use to determine which sources of complaint to tackle first?

(iii) Now suppose the manager has implemented several new procedures, and expects that total claims will be reduced by at least 30%. Set up a control chart to monitor the success of the manager's attempt to reduce the total number of claims. State the assumptions that you have made.

(iv) Now imagine that it is six months since the new procedures were implemented. During this period there have been a total of 28 claims. Construct an approximate 95% confidence interval for the difference in mean monthly total claims between the 20 month before and the six month after period.

F2.   A company buys metal brackets from a supplier in batches of 1000 items. Incoming batches are inspected. A random sample of 40 brackets is taken from each batch. Each bracket in the sample is checked against the specification. If a bracket meets all the specified criteria it is classed as good, but if it fails to meet any of the criteria it is classed as defective. The batch is accepted, as it is, if there is no more than one defective in the sample. If there are two or more defectives in the sample, the entire batch is inspected at the supplier's expense. All defective items are replaced. Assume the proportion of defective items in the batch is p, and that inspection detects all defective items.

(i)   Write down an expression for the approximate probability that the batch is accepted, and evaluate it for p equal to 0.01, 0.05 and 0.10. Why is your answer a very good approximation?

(ii)   Write down an expression for the average outgoing quality (AOQ). By differentiating, or otherwise, find the value of p which corresponds to the maximum AOQ, known as the AOQ limit (AOQL). What is the numerical value of the AOQL? Sketch a graph of the AOQ against p, for values of p from 0 to 0.1.

(iii) The following two-stage sampling scheme has been proposed. Take an initial sample of 15. Accept the batch if no more than one defective is found, reject the batch if 5 or more defectives are found. Otherwise take a second sample of 30. Accept the batch if the total number of defectives in the combined sample is less than 6. If 6 or more defectives are found in the combined sample, reject the batch. If $p = 0.05$, calculate the average sample size inspected, and the probability that a batch will be rejected.

(iv)   What are the disadvantages of acceptance sampling procedures? Suggest an alternative strategy, but state its drawbacks.

F3. The yield of a chemical process is thought to depend on four factors: temperature (A); pressure (B); percentage of catalyst (C); and time (D). The factors can safely be changed by as much as $\pm 10\%$ of their present settings. The plant manager suspects that these settings are not optimum.

(i) Give a suitable half replicate of a full two-level factorial experimental design $\left(2^{4-1}\right)$, and explain clearly how you obtain the half replicate. What are the limitations of this design?

(ii) A regression for the main effects was fitted to the yields recorded in the experiment of part (i). The estimated regression was:

$$y = 43 + 1.2x_1 + 0.9x_2 + 1.5x_3 - 0.4x_4$$

where x_1, x_2, x_3 and x_4 are changes (in units of 2%) from present settings of A, B, C and D respectively. If the temperature (x_1) is increased by 1 unit from present settings, what changes in the other variables correspond to moving in the direction of steepest ascent? What is the predicted yield at this new set point?

(iii) The plant manager wishes to run another, larger, experiment centred on this new set point. One of the objectives is to estimate possible quadratic effects. Three proposals for the experimental design are made at a meeting: run a full factorial design augmented with a star design to make a composite design; run the other half of the half-replicate used in part (i) augmented with a star design; run a one third replicate of a full three level experimental design (3^{4-1}). Discuss, briefly, the advantages and disadvantages of these proposals. Write down the form of a regression model you would fit to the results from the first proposal.

(iv) How might you modify the first proposal in part (iii) if factor C is now "type of catalyst" rather than "percentage of catalyst"? Assume there are only two suitable types of catalyst.

F4.   The parameters α and β in the Weibull distribution, with the cumulative distribution function

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

can be estimated graphically from a Weibull plot. In this plot the natural logarithm of the ith ordered sample value is plotted against $\ln(-\ln(1-(i-0.4)/(n+0.2)))$ where n is the sample size. The gradient of a line drawn through the points is an estimate of $1/\alpha$, and the intercept is an estimate of $\ln\beta$.

The following data are lifetimes (minutes) of windscreen wiper motors moving a blade over a dry screen.

211   223   238   267   275   296   321   341   400   493

(i)   Draw a Weibull plot for these data, and estimate the parameters of the distribution.

(ii)   Subtract 200 from all the data, draw a Weibull plot by adding points to your graph in (i), and estimate the parameters of the distribution.

        [Use a different symbol for the data points in (i) and (ii)].

(iii)   Using your answers to both (i) and (ii) above, estimate the lifetimes that will be exceeded by (a) 99% and (b) 1% of motors. Which of the analyses, (i) or (ii), do you consider gives the more reliable estimates? Justify your answer.

(iv)   A variable X has an exponential distribution with a starting point of L. That is, the cumulative distribution function of X is

$$F(x) = 1 - e^{-\lambda(x-L)}, \qquad \text{for} \quad L \leq x .$$

        Explain why the maximum likelihood estimate of L will always be the sample minimum. Does this argument generalise to the Weibull distribution?