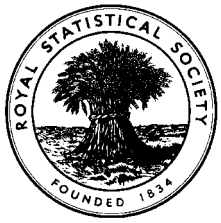
EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



GRADUATE DIPLOMA IN STATISTICS, 2002

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.

Note that $\binom{n}{r}$ *is the same as* ${}^nC_r$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 30 printed pages.
This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2.
Question 1 of Section A starts on page 3.
There are 24 questions altogether in the paper, 4 in each of the 6 Sections.

THIS SECTION STARTS

ON THE

NEXT PAGE

(page 4)

A1. It is proposed to examine the factors influencing Gross Fixed Capital Formation (F) in the United Kingdom. Quarterly data from 1990 to 1999 inclusive (so there are n = 40 observations) in £m at 1995 market prices, not seasonally adjusted, are compiled from Table 1.3 of *Economic Trends Annual Supplement, 2000 edn.*, as are data for Gross Final Expenditure (E). Three quarterly dummies Q_i for i = 2, 3, 4 are defined to equal 1 in the ith quarter and zero otherwise.

Four regressions are calculated as follows, where Δ denotes first differences, and E_{-1} and E_{-2} indicate E lagged one and two quarters respectively, so that $\Delta E \equiv E - E_{-1}$. With conventional notation, with estimated standard errors in brackets,

$$F = \underset{(3199)}{-10139} - \underset{(0.1132)}{0.0992E} - \underset{(0.1559)}{0.0722E_{-1}} + \underset{(0.1184)}{0.3558E_{-2}} - \underset{(2218)}{5496Q_2} + \underset{(2149)}{625Q_3} + \underset{(1881)}{2313Q_4}$$

$$R^2 = 0.918 \quad \bar{R}^2 = 0.903 \quad s = 1250 \quad rss = 51576336 \quad DW = 0.39 \quad \text{……..(a)}$$

$$F = \underset{(2240)}{-7417} + \underset{(0.0097)}{0.1775E_{-2}} - \underset{(569)}{4070Q_2} - \underset{(568)}{743Q_3} + \underset{(567)}{250Q_4}$$

$$R^2 = 0.911 \quad \bar{R}^2 = 0.901 \quad s = 1267 \quad rss = 56203304 \quad DW = 0.26 \quad \text{…….(b)}$$

$$\Delta F = \underset{(0.0415)}{0.2146\Delta E_{-2}} - \underset{(236)}{4191\Delta Q_2} - \underset{(367)}{352\Delta Q_3} - \underset{(318)}{638\Delta Q_4}$$

$$s = 633 \quad rss = 14034007 \quad DW = 2.15 \quad \text{…….(c)}$$

$$\Delta F = \underset{(129)}{-175} + \underset{(0.0526)}{0.2595\Delta E_{-2}} - \underset{(264)}{4359\Delta Q_2} - \underset{(441)}{12\Delta Q_3} + \underset{(383)}{935\Delta Q_4}$$

$$R^2 = 0.925 \quad \bar{R}^2 = 0.916 \quad s = 626 \quad rss = 13310000 \quad DW = 2.17 \quad \text{…….(d)}$$

Question A1 is continued on the next page

(i) Write a commentary, describing why the first regression is unsatisfactory, and why the second, third and fourth regressions were fitted in turn.

(3)

(ii) Which regression is the most satisfactory? Why?

(3)

(iii) The computer program used does not obtain R^2 and $\bar{R}^2$ for regressions which exclude constant terms. Use the results of the fourth regression calculations to obtain R^2 and $\bar{R}^2$ for the third regression.

(3)

(iv) How does the economic model determining F underlying the fourth regression differ from that underlying the third regression? What do you conclude by comparing these two regressions?

(3)

(v) Suppose you wished to examine whether the three quarterly dummies as a set were statistically significant in the third regression. Given the original data, how might you do this?

(3)

(vi) Write an economic account of the determination of F as revealed by these calculations.

(5)

Turn over

A2. In order to examine the relation between household income and consumption per head, data are collected from Table 1.5 of *Economic Trends Annual Supplement, 2000 edn.*, in thousands of pounds at 1995 prices, relating to the n = 20 years 1980 to 1999. Consumption is denoted by C and income by Y. It is found that

$$\sum C = 140.21 \quad \sum Y = 147.24 \quad \sum C^2 = 1005.7963 \quad \sum Y^2 = 1109.9948 \quad \sum CY = 1056.1014$$

With conventional notation, find the simple regression $C = \alpha + \beta Y$, with standard errors of the coefficients. Find r^2 and s for your regression.

(8)

Use your regression to make a prediction of C for the year 2000, conditional on Y equalling 9.40.

(2)

Two 95 per cent intervals, usually called the confidence interval and the prediction interval, are often calculated for such predictions. Obtain them for your prediction, and explain what each of them means.

(4)

If the values of Y and C for 2000 were actually 9.40 and 9.20 respectively, what would you conclude on the basis of your two 95 per cent intervals?

(2)

Test the null hypothesis that the consumption function is really $C = Y$ + error, i.e. $\alpha = 0$ and $\beta = 1$, against the alternative hypothesis that this is not the consumption function.

(4)

A3. (a) Answer the following questions with particular reference to estimating numerical values of economic features of a population.

(i) Explain what is meant by *stratified random sampling* and by *quota sampling*, and distinguish between them.

(3)

(ii) Explain verbally (rather than mathematically) why estimates based on stratified random samples with uniform sampling fractions in each stratum will usually have smaller standard errors than simple random samples of the same size.

(3)

(iii) What are the advantages and disadvantages of stratified random sampling with uniform sampling fractions as compared to quota samples?

(3)

(b) A population of size $N = 10000$ families is divided into three strata, of sizes $N_1 = 5000$, $N_2 = 3000$ and $N_3 = 2000$. Let x_{ij} denote the size of family j in stratum i (for $j = 1, 2, \ldots, N_i$ and $i = 1, 2, 3$). Summary statistics for the stratum populations are as follows.

Stratum	N_i	$\sum_j x_{ij}$	$\sum_j x_{ij}^2$
1	5000	9916	29512
2	3000	7448	25986
3	2000	5983	23733
Total	10000	23347	79231

Find the mean and standard deviation of family size in each stratum and in the population as a whole.

(4)

It is proposed to take a random sample of size $n = 200$ from this population and to use it to estimate the population mean.

Find the standard error of the mean of a simple random sample with replacement of size $n = 200$, and the equivalent standard error of a stratified random sample with replacement of the same total size with a uniform sampling fraction in each stratum.

(4)

Under what circumstances would you expect to find a rather larger difference between standard errors derived from simple random samples and stratified random samples with uniform sampling fractions?

(3)

A4. Investment trusts are companies whose sole activity is owning shares of other companies. Each trust may, typically, hold the shares of 100 or more selected other companies. The dividends paid by investment trusts are thus the dividends which they receive from the companies whose shares they hold, less the cost of running the trusts.

Random samples of 14 investment trusts and 20 other companies are taken and their dividend yields (per cent) noted as follows.

Investment trusts: 0.7 1.3 2.9 3.2 3.3 2.6 2.5 5.0 0.4 2.3 5.4
3.7 2.4 0.0

(sum = 35.7, sum of squares = 123.79)

Other companies: 3.2 0.6 3.7 0.0 3.7 9.3 2.0 3.9 5.5 6.7 4.4
0.0 10.3 5.2 4.4 3.7 6.5 0.0 4.5 4.2

(sum = 81.8, sum of squares = 484.50)

It is proposed to test the null hypothesis that the variances in the two populations are the same. What alternative hypothesis would be most appropriate? Carry out such a test using the above data, making any assumptions necessary for your test.

(6)

Regardless of the result of your test, take it that the variances are in fact equal, and test the null hypothesis that the means in the two populations are the same against an alternative hypothesis to be stated. Provide two possible economic explanations for any possible difference between the two population means.

(6)

On what assumptions does your test of equality of the means depend? Are they realistic?

(4)

Explain, but do not carry out, a non-parametric test which might usefully be applied to these data, specifying carefully the null and alternative hypotheses which it uses.

(4)

SECTION B - ECONOMETRICS

B1. Data from a random sample of 65 metropolitan areas were collected for a study of crime rates. A binary dependent variable, *y*, was defined as follows: $y = 1$ for a high crime rate and $y = 0$ for a low crime rate area. Information was collected on the population size (*S*) in thousands , the population growth rate (*G*) and an education index (*E*).

The probability that area *i* has a high crime rate is modelled by

$$P_i = P(y_i = 1) = F(I_i) = P(Z \leq I_i)$$

where $I_i = \beta_0 + \beta_1 S_i + \beta_2 G_i + \beta_3 E_i$ is the utility index, *F*(.) is the standard Normal cumulative distribution function and *Z* is a standard Normal random variable.

Estimation by maximum likelihood yielded

$$\begin{array}{lcl} \hat{I}_i & = & -1.20 + 0.0010 S_i + 0.056 G_i - 0.40 E_i \\ s.e. & & (0.56)\ \ (0.0009)\ \ (0.022)\ \ (0.15) \end{array}$$

(i) Briefly explain the use of the probit model in econometrics.

(5)

(ii) What are the main differences between the probit and the logit methodologies?

(3)

(iii) Comment on the statistical significance of each parameter and hence interpret the model.

(4)

(iv) What is the implication of the two negative parameters in the estimated model?

(4)

(v) The model was re-estimated, giving the following results:

$$\begin{array}{lcl} \hat{I}_i & = & -1.25 + 0.053 G_i - 0.44 E_i \\ s.e. & & (0.60)\ \ (0.020)\ \ (0.13) \end{array}$$

Two new areas are given to us with the following information on *G* and *E* respectively:

	G	*E*
Area I:	0.10	0.60
Area II:	0.05	0.50

Use the re-estimated model to predict which area is more likely to have a high crime rate. Why do you think using the re-estimated model is more appropriate?

(4)

B2. The demand for beef (Q) is assumed to be determined by the equation

$$\ln Q_t = \beta_1 + \beta_2 \ln PB_t + \beta_3 \ln PL_t + \beta_4 \ln PP_t + \beta_5 \ln Y_t + \varepsilon_t$$

where *PB*, *PL*, *PP* are prices of beef, lamb and pork respectively. *Y* is per capita disposable income. Annual real-price data are available for a particular country for 17 years.

The model was estimated by ordinary least squares using an econometric package, giving the following (edited) output.

```
R-SQUARE = 0.7609            R-SQUARE ADJUSTED = 0.6812
VARIANCE OF THE ESTIMATE-SIGMA**2 = 0.49231E-02
STANDARD ERROR OF THE ESTIMATE-SIGMA = 0.70165E-01
SUM OF SQUARED ERRORS-SSE = 0.59077E-01

SCHWARZ (1978) CRITERION - SC = 0.79959E-02
AKAIKE (1974) INFORMATION CRITERION - AIC = 0.62581E-02

ANALYSIS OF VARIANCE - FROM MEAN
                 SS          DF    MS                F
REGRESSION       0.18798     4.    0.46996E-01       9.546
ERROR            0.59E-01    12.   0.49231E-02       P-VALUE
TOTAL            0.24706     16.   0.15441E-01       0.001

VARIABLE   ESTIMATED         STANDARD     T-RATIO
NAME       COEFFICIENT       ERROR        12 DF       P-VALUE
LOGPB      -0.82657          0.1826      -4.526       0.001
LOGPL       0.19968          0.2127       0.9387      0.366
LOGPP       0.43714          0.3837       1.139       0.277
LOGY        0.10167          0.2940       0.3459      0.735
CONSTANT    4.6726           1.660        2.816       0.016
```

(i) What signs do you expect for each of the coefficients in general? Do you expect this model to be valid for all countries? Why or why not?

(4)

(ii) Interpret the estimated coefficients. Do they seem reasonable?

(4)

(iii) Without any calculation, perform *t* tests on each coefficient and an *F* test for the overall significance of the regression.

(4)

(iv) Suppose you were asked to test whether $\beta_3 = \beta_4 = 0$. You re-estimated the model without these two variables giving the following result.

```
R-SQUARE = 0.6674            R-SQUARE ADJUSTED = 0.6199
VARIANCE OF THE ESTIMATE-SIGMA**2 = 0.58691E-02
```

Test the null hypothesis that the log prices of lamb and pork do not, as a set, have any linear influence on the log demand for beef.

(5)

(v) What extra model would need to be fitted and what analysis carried out to test the hypothesis $\beta_3 = \beta_4$?

(3)

B3. Consider the following simultaneous equation model:

$$\begin{aligned} y_{1t} &+ \beta_{12}y_{2t} + \beta_{13}y_{3t} + \gamma_{11}x_{1t} &&= u_{1t} \\ \beta_{21}y_{1t} &+ y_{2t} + \beta_{23}y_{3t} + \gamma_{22}x_{2t} &&= u_{2t} \\ \beta_{31}y_{1t} &+ y_{3t} + \gamma_{33}x_{3t} &&= u_{3t} \end{aligned}$$

where all endogenous (y) and exogenous (x) variables are measured from their means.

(i) State the criteria for the identifiability of the parameters in an equation of a linear simultaneous equation model.

(5)

(ii) Use these criteria to check the identifiability of each equation in the above model.

(6)

(iii) Some data were collected on these variables. The sums of squares and cross-products were as follows (for example $\mathbf{x_1}'\mathbf{y_2} = \Sigma x_{1t}y_{2t} = 11$).

	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$
$\mathbf{x_1}$	9	–2	–6	5	11	9
$\mathbf{x_2}$	–2	2	4	4	6	–4
$\mathbf{x_3}$	–6	4	10	4	10	–8

Use the data above to estimate the parameters of the third equation in the above model. Comment briefly on your choice of estimator.

You may use the result

$$\begin{bmatrix} 9 & -2 & -6 \\ -2 & 2 & 4 \\ -6 & 4 & 10 \end{bmatrix}^{-1} = \frac{1}{20}\begin{bmatrix} 4 & -4 & 4 \\ -4 & 54 & -24 \\ 4 & -24 & 14 \end{bmatrix}$$

(9)

B4. Write notes on three of the following, including a discussion of their relevance to practical econometric analysis. **(There are 6 or 7 marks for each chosen part.)**

(a) Cointegration and spurious regression.

(b) The Box-Jenkins approach to time-series model building.

(c) Instrumental variables estimation methods.

(d) Selecting the appropriate set of regressors and an appropriate model.

(e) Heteroscedasticity in regression analysis using cross-section data, its causes, and estimation in the presence of such heteroscedasticity.

BLANK PAGE

SECTION C - OPERATIONAL RESEARCH

C1. (i) A project consists of the activities listed in the following table.

Activity	*Prerequisites*	*Duration (weeks)*
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

Draw the project network, and identify the critical path. Show that the shortest time in which the project can be completed is 35 weeks.

(10)

(ii) It is required to reduce the total project duration to 30 weeks. The durations of activities B, G, K and L can be reduced by outsourcing these activities to subcontractors. The reduced durations, and the costs of making a reduction, are shown in the table below. For each activity, the **entire reduction** must be made: it is not possible to make intermediate reductions by partial outsourcing. Find the cheapest way to achieve a reduction to 30 weeks.

Activity	*Reduced duration (weeks)*	*Cost (£000)*
B	3	60
G	7	50
K	1	20
L	6	30

(10)

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

Maximise $-18x_1 + 4x_2 + 5x_3$

Subject to
$$\begin{aligned} 8x_1 - 3x_2 + 3x_3 &\leq 21 \\ 3x_1 + 2x_2 + x_3 &= 6 \\ -2x_1 + 4x_2 + 3x_3 &\geq 15 \end{aligned}$$

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

(10)

(b) A manufacturing company specialising in wrought-iron beds has factories at three sites A, B and C. This month the three factories have produced 15, 18 and 23 beds respectively. These have been sold to four retailers P, Q, R and S, who require 13, 10, 14 and 6 beds respectively. The unit costs in £ of transporting a bed from each factory to each retailer are given in the matrix below. Unsold beds are stored at each factory. Find a transportation scheme which minimises the total cost, and give this minimum total cost.

	P	*Q*	*R*	*S*	*Supply*
A	7	8	13	9	15
B	6	5	15	21	18
C	8	7	6	8	23
Demand	13	10	14	6	

(10)

C3. (a) A car rental company purchases replacement tyres from a supplier. The tyres are used steadily, at a rate of 2000 tyres per year. Storage costs are £20 per tyre per annum.

The tyres can be purchased either from a small local supplier or from a larger national supplier. For the local supplier, the cost of placing an order is £50, and each tyre costs £28.50. For the national supplier, the cost of placing an order is £200 (reflecting the non-local transportation costs), and the purchase price depends on the order size Q as follows:

$$\text{Cost per tyre} = \begin{cases} £28 \text{ for } Q < 300 \\ £27 \text{ for } 300 \le Q < 600 \\ £26 \text{ for } Q \ge 600 \end{cases}$$

By analysing the relevant costs over a long period, determine which supplier should be used, and what order size should be chosen.

(12)

(b) A travel company plans to acquire temporary leases on a small number of luxury apartments for rent in Athens, for the duration of the 2004 Olympic Games. The cost to the company of acquiring each lease is £3500. Up to a week before the start of the Olympics, each apartment can be let for £6000. If any apartments remain unoccupied after this point, the company can rent them out for a reduced price of £2500 each. However, if the company does not acquire sufficient apartments to meet demand, this will result in bad publicity and a potential future loss in business, estimated at £2000 per disappointed client.

The estimated probabilities of demand for the apartments are given in the following table. How many apartments would you advise the company to acquire leases for?

Demand	*Probability*
0	0.00
1	0.05
2	0.10
3	0.15
4	0.15
5	0.20
6	0.20
7	0.10
8	0.05
≥ 9	0.00

(8)

Turn over

C4. (a) Briefly describe two different statistical methods for reducing the variance in the output of a simulation experiment, and describe the main drawbacks of each method.

(6)

(b) A company produces a product on a special machine. This machine is liable to break down, and it then has to be repaired before it can be used again. Every time the machine breaks down, the cost to the company is £2000 per day in lost production. For the past three years, the company has found that the total cost of these breakdowns has been over £80 000 per annum.

The company is considering introducing a maintenance programme in order to reduce the total machine downtime. The maintenance programme would cost £20 000 per annum, but would result in a longer time between breakdowns. The engineering department estimates that the distribution of the time between breakdowns (i.e. from the end of a repair until the next breakdown) would be defined by the following continuous p.d.f.

$$f(x) = \frac{x}{18} \quad \text{for } 0 \le x \le 6 \text{ weeks.}$$

Under the maintenance programme, the repair time would also tend to be reduced, and the engineers estimate it would have the following discrete distribution.

Repair time in days	Probability
1 (= 0.14 weeks)	0.4
2 (= 0.29 weeks)	0.5
3 (= 0.43 weeks)	0.1

Using the following two streams of random numbers, simulate the system with the maintenance programme for one year (i.e. until the cumulative time is greater than 52 weeks). Work to 2 decimal places.

Random numbers for the times between breakdowns: 0.45, 0.90, 0.84, 0.17, 0.74, 0.94, 0.07, 0.15, 0.04, 0.31, 0.07, 0.99, 0.97.

Random numbers for the repair times: 0.19, 0.65, 0.51, 0.17, 0.63, 0.85, 0.37, 0.89, 0.76, 0.71, 0.34, 0.11, 0.27.

On the basis of this simulation, would you advise the company to go ahead with the programme? What cautionary advice would you give the company before they adopted your recommendation?

(14)

SECTION D - MEDICAL STATISTICS

D1. (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) Acute lower respiratory tract infection is one of the commonest causes of death among infants and children under 5 in developing countries.

When an infant has acute respiratory infection, it is important to decide whether the infant has lower respiratory tract infection (LRI) and should receive antibiotics, or whether the infant has upper respiratory tract infection (URI). The following data come from a study of the usefulness of the respiratory rate for this purpose in infants.

Respiratory rate (breaths/min)	**Number of children (% of total)**			
	LRI		*URI*	
0 – 30	1	(1%)	16	(11%)
31 – 40	4	(3%)	77	(51%)
41 – 50	10	(7%)	46	(30%)
51 – 60	41	(29%)	9	(6%)
61+	86	(61%)	3	(2%)
Total	142	(100%)	151	(100%)

Source: Cherian, T., et al. Lancet 1988.

(i) Four possible cut-off values for a diagnostic test are 30, 40, 50 and 60 breaths/min. Determine the corresponding test sensitivities and specificities.

(4)

(ii) Sketch the ROC curve using the four cut-off values from part (i). Which cut-off gives the best balance between sensitivity and specificity?

(6)

(iii) The authors of the report estimated that the prevalence of LRI among all infants with acute respiratory infection in a developing country is 3%. Calculate the positive and negative predictive values when the prevalence is 3%, and discuss the relative advantages and disadvantages of using a cut-off point of 50 or 60 when the prevalence is 3%.

(6)

D2. The survival times from diagnosis (in months) for a random sample of 24 patients with colorectal cancer from one centre are given below.

3*, 6, 6, 6, 6, 8, 8, 12, 12, 12*, 15*, 16*, 18*, 18*, 20, 22*, 24, 28*, 28*, 28*, 30, 30*, 33*, 42 (* Indicates a right-censored observation.)

Source: McIllmurray, M.B., and Turkie, W. BMJ 1987.

(i) Explain what is meant by a *right-censored observation*.

(2)

(ii) Compute the Kaplan-Meier estimate of the survival curve and plot it. Using Greenwood's formula, calculate the associated standard error for the Kaplan-Meier survival function estimate at 12 months follow-up.

(12)

(iii) Find an approximate 95% confidence interval for the one-year survival rate for colorectal cancer patients from this centre.

(4)

(iv) Use your graph to estimate the median survival time for colorectal cancer patients from this centre.

(1)

(v) The 24 patients were part of a larger randomised controlled clinical trial to compare two treatments, control and γ-linolenic acid, for colorectal cancer. If the data from all patients in the study were available, give the name of an analysis which could be used to compare the survival times for these two treatments.

(1)

D3. A public health authority is planning an epidemiological survey to estimate the prevalence of angina in the adult population in the area the authority covers. The health authority is intending to survey, by means of a postal questionnaire, a random sample of adults in the population.

The director of public health asks you as the study statistician: "how many subjects do we need to respond to the survey in order to assess the prevalence of angina with a reasonable degree of accuracy?"

(i) What is the difference between the incidence and prevalence of a disease?

(2)

(ii) A sample estimate p of the true prevalence π of angina is available, based on a previous survey. A 95% confidence interval is now required for the true value of π in the population. This interval is to be of width 2ε; i.e. if $\hat{\pi}$ is the new estimate of π, the lower and upper limits of the interval will be $\hat{\pi}-\varepsilon$ and $\hat{\pi}+\varepsilon$. Derive an approximate formula for the necessary number, N, of respondents to the planned survey in order to achieve this accuracy.

(8)

(iii) From the previous survey of 1400 subjects, 20 years ago, the estimated proportion of angina sufferers was p = 0.15 (210/1400). The director of public health thinks the current proportion of angina sufferers is still around 0.15. How many subjects need to respond to the planned survey if we wish to determine this prevalence to within ±0.015, that is, to have a 95% confidence interval with a total width of 0.03?

(2)

(iv) The survey is carried out and 1500 subjects respond to the postal questionnaire. The estimated prevalence of angina is now 20% (300/1500). Has the prevalence of angina changed over the 20-year period between the first and second survey? Perform an appropriate hypothesis test and comment on the results.

(6)

(v) Comment on the use of a postal questionnaire for such a survey.

(2)

D4. (a) Discuss the use of controls, placebos and blinding in clinical trials.

(8)

(b) Sixty-five pregnant women at a high risk of pregnancy-induced hypertension participated in a randomised-controlled trial comparing 100mg of aspirin daily and a matching placebo during the third trimester of pregnancy. The observed rates of hypertension are shown in the table below.

Hypertension rates in aspirin and placebo groups

	Aspirin treated	*Placebo treated*	*Total*
Hypertension	4	11	15
No hypertension	30	20	50
Total	34	31	65

Source: Schiff, E., et al. N. Engl. J. Med. 1989.

(i) Do these data suggest that daily aspirin reduces the risks of hypertension in the last trimester of pregnancy? Perform a chi-squared test with Yates' continuity correction to compare the hypertension rates in the aspirin and placebo groups. Comment on the results of this hypothesis test.

(6)

(ii) Calculate a 95% confidence interval for the difference in hypertension rates between the aspirin and placebo groups. Does the confidence interval estimate from these data suggest that daily aspirin reduces the risk of hypertension in the last trimester of pregnancy?

(6)

SECTION E - BIOMETRY

E1. In a field experiment designed in randomised (complete) blocks, one observation is missing and there is some doubt about the accuracy of another. The available data, on crop yield in the early part of a season, are given in kg in the following table; the figure in brackets is of doubtful accuracy and * indicates the plot where the observation is missing.

	Treatment				
Block	*A*	*B*	*C*	*D*	*E*
I	14.2	*	19.3	18.4	19.5
II	15.1	14.3	18.9	16.2	17.9
III	14.4	12.2	16.1	15.3	18.4
IV	16.3	17.0	19.8	16.0	(12.1)

(i) Derive a general formula which can be used to estimate a single missing value in a randomised block design containing t treatments and b blocks, where the 'incomplete' totals in the block and treatment affected are B', T' respectively, and the 'incomplete' grand total is G'.

Explain briefly why the residual on the affected plot will be zero.

(6)

(ii) Apply this formula to find an estimate of * in the data above. For this purpose you may use the value 12.1 (which is in doubt).

(2)

(iii) The experimenter suspects a recording error in the value (12.1) and wants this to be replaced by another estimated figure. Show how the method in part (i) can be extended to estimate two values simultaneously, and carry out that estimation for the data above. If you recommend an iterative method, only one iteration need be carried out.

(5)

(iv) Another method of dealing with missing observations is to use an analysis of covariance.

(a) Explain what advantages this has over the 'formula' method when the full data analysis is carried out and treatment means are being compared.

(4)

(b) *Without carrying out any further calculation*, explain what you would use as covariates to analyse this experiment with both missing and doubtful observations omitted, and what linear model would form the basis of a computer analysis.

(3)

E2. The following data, expressed in suitable units, were obtained in an experiment to study bean root cells, whose water content y was measured at various distances x from the growing tip.

y	1.3	1.3	1.9	3.4	5.3	7.1	10.6	16.0	16.4	18.3	20.9	20.5	21.3	21.2	20.9
x	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5

(i) Plot the data.

(4)

(ii) The researcher wished to fit a non-linear regression model to the data using the method of least squares. It was decided to use a general Gompertz model

$$y = A + C\exp\left(-e^{-B(x-M)}\right)$$

in which A, B (>0), C and M are parameters to be estimated.

(a) Discuss briefly the computing problems in obtaining least squares estimates of the parameters. Suggest an alternative approach to a solution, based on linear regression and using information from the graph.

(4)

(b) Comment on the assumptions made about the variance of y in your suggested approach in (a), compared with those in the researcher's original proposal.

(2)

question E2 continued on next page

(iii) Computer output from fitting this model gave:

```
Fitted Curve:  Y = A + C*(EXP(-EXP(-B*(X-M))))

Constraints:   B > 0.


SUMMARY OF ANALYSIS

                  DF           SS           MS
 Regression        3         943.084      314.3614
 Residual         11           6.845        0.6223
 TOTAL            14         949.929

Percentage variance accounted for 99.1
Message: Unit 8 has large residual, value 2.15


ESTIMATES OF PARAMETERS

              Estimate     S.E.
        B       0.4785     0.0523
        M       5.819      0.169
        C      20.205      0.794
        A       1.651      0.464
```

An extract from the output of fitted values of Y was:

x	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5
$\hat{y}$	1.65	1.66	1.80	2.62		7.95	11.47		16.96	18.67	19.82	20.57	21.05		21.54

(a) Complete the table of fitted values of Y.

(b) Plot the fitted curve on the same graph as the data, and comment on the fit.

(c) Explain the meaning of the "Message" in the output.

(7)

(iv) Suggest what alternative models might have been studied for explaining these data. Would you expect any to be better than the Gompertz model, and, if so, why?

(3)

E3. (i) Suppose that a ratio of two parameters α and β is to be estimated by a/b, where a and b are unbiased estimators of α, β and each is a linear function of a set of observations with Normally distributed residual errors. An analysis of variance has provided an estimate, s^2, of the variance of the observations, with f degrees of freedom. Using this, the estimated variances and covariance of a and b have been calculated as s^2v_{11}, s^2v_{22} and s^2v_{12} (where v_{11}, $v_{22,}$ v_{12} depend on the definitions of a and b).

(a) Prove the result known as Fieller's Theorem, that confidence limits for the true value of the ratio $\mu = \alpha / \beta$ are

$$\frac{m - \frac{gv_{12}}{v_{22}} \pm \frac{ts}{b}\left\{v_{11} - 2mv_{12} + m^2v_{22} - g\left(v_{11} - \frac{v_{12}^{\ 2}}{v_{22}}\right)\right\}^{1/2}}{1-g},$$

in which $m = a/b$, t is the critical value of Student's t, on f degrees of freedom, at the required probability level, and $g = t^2s^2v_{22}/b^2$.

(b) Under what conditions can g be neglected?

(10)

(ii) A set of experimental data conforms to a model in which a linear regression $y = c + dx$ holds from $x = 0$ to $x = X$ and another linear regression $y = \gamma + \delta x$ holds for all $x \geq X$. The exact value of X is not known, but the first n_1 data points are known to have values of x which are less than X and the remaining n_2 data points have values of x which are greater than X.

(a) Assuming the usual properties required for regression analysis, show how to estimate X from the data.

(b) Justify the use of Fieller's Theorem for obtaining confidence limits for the true value of X, and explain how to obtain s^2, f, v_{11}, v_{22} and v_{12} in this case.

[You may use the following results from simple linear regression with the model $y = A + Bx + e$, where e is a Normally distributed residual (error) term with constant variance σ^2:

(1) $\mathrm{Var}\left(\hat{A}\right) = \sigma^2 \sum x^2 / \left(nS_{xx}\right)$,

(2) $\mathrm{Var}\left(\hat{B}\right) = \sigma^2 / S_{xx}$,

(3) $\mathrm{Cov}\left(\hat{A}, \hat{B}\right) = -\bar{x}\sigma^2 / S_{xx}$,

and $S_{xx} = \sum_i \left(x_i - \bar{x}\right)^2$.]

(10)

E4. ANSWER ANY TWO OF (a) – (e).

(There are 10 marks for each chosen part.)

Your answers should be specifically on agricultural and biometric aspects of each topic, and should be illustrated by appropriate examples. General answers, without this specific application, will gain very few marks.

(a) Discuss the design of questionnaires for use in agricultural surveys, with reference to examples known to you from your own country.

(b) Explain methods of estimating crop yield in a field experiment by sampling at harvest time, illustrating by reference to any crop known to you. Mention the strengths and weaknesses of any methods you discuss.

(c) Discuss the problems in planning and carrying out experiments on different field sites OR involving different laboratories. Your answer should include discussion of problems (such as variance not constant over sites) that will affect the analysis.

(d) Sampling frames may not be available in suitable form when planning an agricultural survey, particularly in developing regions and isolated parts of a region. Discuss the construction of frames, using aerial surveys and land maps, and explain how a sampling method (e.g. cluster, multi-stage) might be chosen for such a survey.

(e) Discuss the special characteristics and methods of analysis of long-term experiments (those which last several seasons), with special reference to any crop known to you.

BLANK PAGE

SECTION F - STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. (i) Let X_1 and X_2 be independent random variables from a Normal distribution with mean μ and standard deviation σ. What is the distribution of $(X_1 - X_2)$? Hence, or otherwise, deduce that the median of the random variable

$$\left|(X_1 - X_2)\right|$$

is 0.955σ.

(5)

(ii) The specification for the gain on a semiconductor device is that it should be between 260 and 340. The following 21 observations are measurements of gain for individual devices, taken from the production line (in the order shown) at approximately hourly intervals over three successive shifts while the process was operating continuously.

Shift 1 (morning)	291	290	308	300	284	297	298
Shift 2 (afternoon)	306	310	302	304	306	318	314
Shift 3 (night)	276	270	276	274	264	269	262

The mean ($\bar{x}$) and the standard deviation (s) of these 21 observations are 291.4 and 17.5 respectively.

(a) Plot a runs chart for the 21 observations showing the mean.

(3)

(b) Calculate the moving ranges of the 21 observations, and then calculate the median of these 20 ranges. Use the result from part (i) of this question to obtain an estimate ($\hat{\sigma}$) of the process standard deviation. Add lines on your chart at $\bar{x} \pm 3\hat{\sigma}$, and comment on the result.

[Note. A moving range is the absolute value of the difference between two consecutive observations. Thus the first two moving ranges for the above data are 1 and 18.]

(7)

(c) Compare $\hat{\sigma}$ with s, and comment. How would you summarise the process capability if the current situation is allowed to continue? What process capability might be achieved? How would you attempt to achieve this?

(5)

F2. (a) A company manufactures a capacitor with a target capacitance of 68 microfarad (μF). The process standard deviation is known to be 2.4 μF from extensive past records. The process is monitored by measuring the capacitances of random samples of 5 capacitors every hour.

(i) Describe, with the aid of a diagram, how you would set up a Shewhart mean chart and the associated range chart. [You may assume that the factors for the standard deviation when adding upper and lower action lines on a range chart for samples of five are 0.37 and 5.48.]

(8)

(ii) Demonstrate the use of your chart with the following data.

Sample number		
1	*2*	*3*
68.71	70.06	71.10
64.79	71.01	68.60
69.71	64.18	70.86
71.90	64.54	75.95
66.34	69.37	73.45

(2)

(iii) Suppose the mean shifts to 70 μF. What is the probability that the next sample mean lies beyond the upper action limit? What is the expected number of hours before action is indicated?

(3)

(b) The manager of a large bakery keeps records of customer complaints. These averaged 0.14 per day until new ovens were installed six months ago. Since then, the numbers of complaints in six 30-day periods have been 7, 5, 6, 3, 11 and 8. Set up a suitable control chart, and plot these data. Do you think there is evidence of an increase in complaints since the installation of the new ovens?

(7)

F3. A company's business is to repair photocopiers. There is one repair man in town *A* in which there are many machines. Call-outs occur according to a Poisson process with an average rate of λ per day. The man can repair machines at an average rate of θ per day ($\theta > \lambda$), and repair times have an exponential distribution.

Take the state space to be the number of machines in the system, that is the number of machines awaiting repair and under repair.

(i) Write down the equations for the steady state situation.

(4)

(ii) Solve the equations in (i) to obtain expressions for the proportions of time there are 0, 1, 2, … machines in the system, in terms of λ and θ.

(4)

(iii) Show that the mean number of machines in the system is $\lambda/(\theta-\lambda)$.

(5)

(iv) Deduce from the result in (iii) an expression for the average waiting time before service begins.

(2)

(v) Write down equations for the steady state situation if there are two repair men. Assume only one man can work, at any one time, on a photocopier that needs repairing, and the repair times for both men have the distribution described above.

(5)

F4. An experiment was carried out with the objective of reducing variability of the epitaxial layer on integrated-circuit wafers. The target thickness was 20 microns. Six factors were investigated, each at two levels according to an L_8 array: arsenic gas flow rate (A), deposition temperature (B), deposition time (C), nozzle height (D), acid flow rate (E), and acid etch temperature (F). For each combination of factors, the thickness of the epitaxial layer was measured on 25 wafers. The means and the natural logarithms of the standard deviations for each set of 25 observations follow.

A	B	C	D	E	F	(G)	$\bar{y}$	ln(s)
–	–	–	–	–	–	–	19.8	0.73
–	–	–	+	+	+	+	18.8	0.48
–	+	+	–	–	+	+	21.5	0.93
–	+	+	+	+	–	–	20.9	1.09
+	–	+	–	+	–	+	21.0	0.10
+	–	+	+	–	+	–	20.2	0.60
+	+	–	–	+	+	–	20.0	0.85
+	+	–	+	–	–	+	18.7	0.62

(i) (a) The first six columns of the L_8 array are equivalent to a fractional factorial design. Which is it?

(b) Explain what is meant by aliasing, and define the resolution of a design. What is the resolution of this design? Identify a two-factor interaction which is an alias of the main effect of A.

(4)

(ii) The spare column (G) is aliased with three two-factor interactions. Write down any **one** of them.

(2)

(iii) A regression of $\bar{y}$ on the six factors, each coded ±1, gave

$$\bar{y} = 20.112 - 0.137A + 0.163B + 0.788C - 0.463D + 0.062E + 0.013F$$

with a standard deviation of residual errors equal to 0.3182 on 1 degree of freedom. Construct a 90% confidence interval for the coefficient of C.

(4)

(iv) If (G) is included in the regression of part (iii), its coefficient is –0.133. Why are the other coefficients unchanged? Draw a half-Normal plot of the seven coefficients using the plotting points on the horizontal axis: 0.088, 0.265, 0.452, 0.656, 0.893, 1.194, 1.680.

(3)

(v) A regression of ln(s) on the six factors gave

$$\ln(s) = 0.675 - 0.133A + 0.198B + 0.005C + 0.022D - 0.045E + 0.040F\ .$$

Now include (G) in the regression and calculate its coefficient. Draw a half-Normal plot of the seven coefficients.

(4)

(vi) What practical conclusions would you draw from your analysis, and on what assumptions do these depend? How would you advise the company to proceed?

(3)