EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 2001

Options Paper

Time Allowed: Three Hours

This paper contains four questions from each of six option syllabuses. Each option syllabus is one Section.

Section A: Statistics for Economics
B: Econometrics
C: Operational Research
D: Medical Statistics
E: Biometry
F: Statistics for Industry and Quality Improvement

Candidates should answer **FIVE** *questions chosen from* **TWO SECTIONS ONLY**.

Do **NOT** *answer more than* **THREE** *questions from any* **ONE** *Section.*

ANSWER EACH SECTION IN A SEPARATE ANSWER-BOOK.

Label each book clearly with its Section letter and name.

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

Where a calculator is used the method of calculation should be stated in full.

This examination paper consists of 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)

Turn over

A1. In order to investigate the demand for imports to the United Kingdom, quarterly data (£m, constant 1995 prices, seasonally adjusted) for the period 1985 to 1998 inclusive were collected from *Economic Trends Annual Supplement* 1999 edition, Table 1.3, relating to total imports (M) and gross final expenditure (C). A time trend t takes the values 0, 0.25, …, 13.75 over the $n = 56$ observations. The following ordinary least squares regressions were obtained, where $m = \ln M$ and $c = \ln C$, with estimated standard errors in parentheses:

$$\underset{(0.2438)}{m=-11.3067}+\underset{(0.0199)}{1.7963c} \qquad r^2=0.993 \qquad rss=0.0158 \qquad s=0.01713 \qquad \text{(A)}$$

$$\underset{(0.0143)}{m=10.3771}+\underset{(0.0018)}{0.0496t} \qquad r^2=0.934 \qquad rss=0.1593 \qquad s=0.05431 \qquad \text{(B)}$$

$$\underset{(0.9330)}{m=-10.3182}+\underset{(0.0773)}{1.7143c}+\underset{(0.0022)}{0.0024t} \qquad r^2=0.994 \qquad rss=0.0155 \qquad s=0.01710 \qquad \text{(C)}$$

$$\underset{(0.9457)}{m=-10.2496}+\underset{(0.0783)}{1.7086c}+\underset{(0.000179)}{0.000217(ct)} \qquad r^2=0.994 \qquad rss=0.0155 \qquad s=0.01707 \qquad \text{(D)}$$

(i) Give a brief economic interpretation of the values of the coefficients of the variables in the above four equations.

(4)

(ii) Account for the differences between the coefficients of c and t in the above four equations.

(4)

(iii) Test the null hypothesis that, in the model $m = \beta_0 + \beta_1 c + \beta_2 t + u$, where u is a stochastic term with the usual properties, β_1 is 1.000.

(4)

question A1 continued on next page

The model $m = \gamma_0 + \gamma_1 c + \gamma_2 (ct) + u$, being the one with the smallest s statistic, is fitted separately to the five-year periods 1985-1989 and 1994-1998, with the following results respectively:

$$\underset{(4.294)}{m = -15.825} + \underset{(0.3569)}{2.1723c} - \underset{(0.001458)}{0.001791(ct)} \quad r^2 = 0.986 \quad rss = 0.00457 \quad s = 0.01639 \quad \text{(E)}$$

$$\underset{(6.992)}{m = -12.714} + \underset{(0.5851)}{1.9064c} + \underset{(0.001818)}{0.000281(ct)} \quad r^2 = 0.913 \quad rss = 0.00194 \quad s = 0.01067 \quad \text{(F)}$$

(iv) Test the stochastic term for homoscedasticity.

(4)

(v) Use large sample methods to test the null hypothesis that the coefficient of c was the same in the two five-year periods, the difference between 2.1723 and 1.9064 shown above reflecting sampling. If one had access to the original data, how otherwise might one test this null hypothesis?

(4)

A2. In order to investigate the market for the shares of retail companies, statistics of the yields (percent) of the shares in 13 food retailing companies and 20 other retailers were compiled from *The Times* newspaper of 19 May 2000 as follows:

Food retailers	0.0	3.0	8.6	2.9	2.0	2.2	1.0	3.7	4.1	19.6
	2.7	2.2	6.0							
	(sum 58.0, sum of squares 564.00)									
Other retailers	6.6	7.9	6.5	4.4	5.0	1.6	5.4	2.8	7.9	10.2
	8.6	1.9	0.0	2.1	1.1	4.6	0.0	1.8	3.4	2.8
	(sum 84.6, sum of squares 525.58)									

You may treat the two sets of data as simple random samples from large populations.

Test the null hypothesis that the variances in the two populations were the same.

(4)

The data are analysed using the Minitab statistics package as shown on the facing page of this examination paper, the data having been entered into C1 and C2 respectively. Explain what analyses have been carried out, and any relationships among them.

(6)

What assumptions, if any, are the Minitab analyses based upon? How realistic are they? (You should consider both the actual data and the economic conditions that generated them.)

(5)

Write a paragraph of advice to a prospective investor on the advantages and disadvantages of investing in food retailers as compared with other retailers, using your results and indicating what further information you would like to make use of.

(5)

Minitab output for this question is on next page

```
MTB > twosample c1 c2

Two sample T for C1 vs C2

     N      Mean     StDev   SE Mean
C1  13      4.46      5.04       1.4
C2  20      4.23      2.97      0.66

95% CI for mu C1 - mu C2: ( -3.0,  3.50)
T-Test mu C1 = mu C2 (vs not =): T = 0.15  P = 0.88  DF = 17

MTB > twosample c1 c2;
SUBC> pool.

Two sample T for C1 vs C2

     N      Mean     StDev   SE Mean
C1  13      4.46      5.04       1.4
C2  20      4.23      2.97      0.66

95% CI for mu C1 - mu C2: ( -2.6,  3.07)
T-Test mu C1 = mu C2 (vs not =): T = 0.17  P = 0.87  DF = 31
Both use Pooled StDev = 3.91

MTB > aovoneway c1 c2

Analysis of Variance
Source     DF        SS        MS        F        P
Factor      1       0.4       0.4     0.03    0.869
Error      31     473.0      15.3
Total      32     473.4
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  --+---------+---------+---------+----
C1         13     4.462     5.043   (-----------------*------------------)
C2         20     4.230     2.971    (--------------*--------------)
                                   --+---------+---------+---------+----
Pooled StDev =    3.906            2.4       3.6       4.8       6.0

MTB > mann-whitney c1 c2

C1         N =  13     Median =       2.900
C2         N =  20     Median =       3.900
Point estimate for ETA1-ETA2 is      -0.500
95.1 Percent CI for ETA1-ETA2 is (-2.700,1.402)
W = 211.5
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.7402
The test is significant at 0.7400 (adjusted for ties)

Cannot reject at alpha = 0.05
```

A3. It is often claimed that a major weakness in the UK economic structure is the low proportion of Gross Domestic Product (GDP) which is invested. The then government endeavoured to raise this proportion in the two or three years following the General Election of 1987 by stimulating a greater rate of economic growth ("the Lawson boom"). In order to examine this problem, annual data are collected from *United Kingdom National Accounts* 1999 edition (the Blue Book), Table 1.3, covering the 18 years 1981 to 1998, from which *I*, the percentage of gross investment in GDP, is calculated with the following results:

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989
I	13.49	14.61	15.50	16.45	16.41	16.10	16.82	18.92	19.18
	1990	1991	1992	1993	1994	1995	1996	1997	1998
	17.90	16.11	16.48	16.55	17.03	16.95	16.93	17.85	19.18

A variable t is defined to take the values −8.5, −7.5, …, 8.5. It may be found that
$\sum t = 0,\ \sum t^2 = 484.50,\ \sum I = 302.46,\ \sum I^2 = 5119.4122,\ \sum tI = 83.370$.

Find the regression of I on t, calculate r^2 and s and estimate the standard errors of the coefficients in the regression.

(4)

Use your regression to predict I in the year 2001.

(2)

Two 95 percent intervals are often compiled for such regression predictions, frequently called "confidence interval" and "prediction interval" (e.g. in the output of standard computer statistics packages). Calculate them for this example and explain to what they refer.

(4)

Draw a time chart of the original data above, and draw your regression line on it.

(4)

Discuss in detail the light your analyses throw on the problem of alleged UK under-investment.

(6)

A4. Give and explain formulae for Laspeyres (base-weighted) price index numbers, in terms of (i) costs of baskets of goods etc, (ii) weighted averages of price relatives, and prove that the two formulae give the same results.

(4)

As far as is practicable, give and explain similar formulae for Paasche (current-weighted) price index numbers.

(2)

In times of rising prices, which type of index number would be expected to show the larger rate of price increase? Why?

(2)

Which kind of index number is more commonly used in practice? Why?

(2)

UK exports of goods and services, 1990 - 1998, £bn

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998
At 1995 prices	156.6	156.4	162.8	169.2	184.9	202.4	217.6	236.3	241.1
At current prices	133.5	135.4	143.2	162.1	178.8	202.4	220.3	229.3	224.2

(Source: *United Kingdom National Accounts, 1999 edition, Tables 1.2 and 1.3*)

Using the above data, compile a series of an index of price of exports and a series of an index of quantity of exports. What kind(s) of index numbers are they?

(6)

What do your results show?

(4)

BLANK PAGE

SECTION B - ECONOMETRICS

B1. An economist wants to estimate, on the basis of quarterly data, a model relating savings, y_t, to disposable income, x_t, and postulates the regression

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

where ε_t is a non-autocorrelated zero-mean random error term, uncorrelated with x_t. Although data are available through attempts to measure savings and disposable income, the economist suspects that these observations contain random measurement errors.

(i) Analyse the impact on the least squares estimators of (α, β) of random measurement error only in savings.

(5)

(ii) Analyse the impact on the least squares estimators of (α, β) of random measurement error only in disposable income.

(8)

(iii) Discuss the use of instrumental variables to overcome any shortcomings of least squares estimators caused by measurement errors of either or both of the above types.

(7)

B2. (i) What is meant by saying that a time series is generated by an ARIMA(p,d,q) model? Describe a methodology for fitting such a model to data.

(10)

(ii) Outline a formal test of the null hypothesis that the appropriate degree of differencing, d, in an ARIMA model for a given series is 1 against the alternative hypothesis $d = 0$. Discuss any difficulties that might be encountered in the practical application of the test.

(10)

B3. (i) What is meant by saying that the error terms in a multiple regression model exhibit heteroscedasticity, and what are the consequences if heteroscedasticity is ignored and the usual ordinary least squares estimation and inference procedures are applied?

(7)

(ii) Why is it desirable to apply more than one test for heteroscedasticity? Outline at least two alternative test procedures.

(7)

(iii) Discuss feasible estimation procedures that might be applied in the presence of particular types of heteroscedasticity.

(6)

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

(a) Logit analysis.

(b) The identification problem in a simultaneous system of econometric equations.

(c) Testing for autocorrelation in the errors of a regression with a lagged dependent variable.

(d) Cointegration and its implications.

(e) Multicollinearity.

SECTION C - OPERATIONAL RESEARCH

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

Activity	*Prerequisites*	*Shortest duration*	*Most likely duration*	*Longest duration*
A	-	8	10	12
B	-	6	14	24
C	A	1	3	5
D	A	7	8	9
E	B	2	4	16
F	C	3	5	7
G	B, D	6	8	10
H	F, G	5	6	9
I	E	2	12	18

(i) Draw the project network, and identify the critical path using the expected activity durations.

(6)

(ii) Making any assumptions needed, estimate the probability that the project will be completed within 35 days. State the assumptions you make clearly.

(10)

(iii) Are these assumptions justified in general – and in this particular case?

(2)

(iv) How could you improve your estimate?

(2)

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

$$\text{Maximise } 3x_1 + 2x_2 - 7x_3 + x_4$$

$$\begin{aligned} \text{Subject to} \quad 4x_1 + 5x_2 + 2x_3 \quad &\leq 9 \\ x_1 \qquad\qquad + x_4 &= 3 \\ x_1 - x_2 + 2x_3 \quad &\geq 5 \\ \\ x_1, x_2, x_3, x_4 &\geq 0 \end{aligned}$$

(8)

If the right hand side of the "≤" constraint changes to 9 + δ, for what range of values of δ would the change in the optimal objective value be proportional to δ?

(2)

(b) A greenhouse manufacturer has factories at three sites *A*, *B* and *C*. This month these three factories have produced 30, 50 and 80 greenhouses respectively, and these have been sold to four garden stores *P*, *Q*, *R* and *S*, who require 20, 40, 50 and 50 greenhouses respectively. The unit costs in £ of transporting a greenhouse from each factory to each garden store are given in the matrix below. Find a transportation scheme which minimises the total cost.

	P	*Q*	*R*	*S*	*Supply*
A	25	21	25	14	30
B	15	10	18	24	50
C	24	20	18	13	80
Demand	20	40	50	50	

(10)

C3. (a) The notation *M*/*M*/*s* denotes a queueing system with *s* servers, where arrivals are random and independent and service times are exponentially distributed.

(i) Show that the average time in the system for the *M*/*M*/1 queue with mean arrival rate λ and mean service time $1/\mu$ is $W_1 = 1/(\mu - \lambda)$.

(ii) Show that W_1 is always less than the average time in the system for the *M*/*M*/2 queue with mean arrival rate λ and mean service time $2/\mu$.

[Note: For an *M*/*M*/2 queue with mean service time $1/\nu$, it can be shown that the average time in the system is given by $W = 4\nu/(4\nu^2 - \lambda^2)$.]

(10)

(b) A petrol station has two petrol pumps and space for another three cars to queue. Customers arrive at random at a mean rate of one every minute. Customers who arrive to find no queueing space leave and do not return. Service times are exponentially distributed with a mean of 3 minutes. It is proposed that a third petrol pump be installed, leaving space for 2 cars to queue. The petrol station is open for 40 hours per week. If the average profit per customer is 50p and the third pump would cost £150 per week, should the third pump be installed? (You may ignore the cost of installing the third pump.)

(10)

C4. Company *A* manufactures tractors, but purchases their engines from a supplier. Demand for the engines is steady and the annual demand is 3000 engines. The cost of placing an order for a replenishment is £7500. The company owns a facility that can store up to 200 engines at a cost of £80 per engine per annum. To obtain additional storage space, the company also has the option of leasing one of the following warehouses: a small warehouse, which costs £50000 per annum, with a capacity of 300 engines, or a large warehouse, which costs £60000 per annum, with a capacity of 600 engines. Storage costs in each of the leased warehouses are also £80 per engine per annum.

Should the company lease either of the warehouses, and what order quantity for engines should be used?

(12)

Company *A* decides to lease the small warehouse. It is approached by Company *B*, which wishes to rent *A*'s own storage facility for 200 engines. What is the minimum annual rental which *A* should accept?

(8)

SECTION D - MEDICAL STATISTICS

D1. What is the difference between a *parallel group design* and a *cross-over design* in a clinical trial to compare two treatments?

(2)

In a trial of a new drug for the treatment of asthma, each of 10 patients was given the drug for a period of 14 days and a placebo for a separate period of 14 days, the order of administration being chosen randomly for each patient. The table shows the maximum % fall in FEV1 on the drug and placebo after six minutes on an exercise treadmill following treatment.

Group A (drug / placebo)			**Group B (placebo / drug)**		
Patient	*Period 1*	*Period 2*	*Patient*	*Period 1*	*Period 2*
2	18.70	8.47	1	20.00	8.47
3	48.89	20.45	4	22.98	0.12
6	2.99	20.29	5	26.47	10.53
7	17.07	23.30	8	24.84	7.28
9	16.05	36.10	10	20.66	10.94
Mean	20.74	21.72	*Mean*	22.99	7.47
Standard deviation	16.93	9.86	*Standard deviation*	2.73	4.37

Stating any assumptions you make, test for (i) a period effect, (ii) a treatment × period interaction, (iii) a treatment effect, and report on your conclusions.

(18)

D2. The table below shows the age distribution of male deaths from motor vehicle traffic accidents in England and Wales in 1996. It also shows the corresponding age distributions for the male England and Wales population and the European Standard Population in 1996.

(i) Explain why standard populations are useful when comparing the mortality of several groups.

(3)

(ii) Calculate the age-specific male mortality rates for motor vehicle traffic accidents separately for England and for Wales.

(7)

(iii) What is unusual about these age-specific mortality rates?

(1)

(iv) Calculate a direct age-standardised mortality rate (using the European Standard Population) separately for each of the two countries, and comment on the relative mortality for the two countries.

(9)

Motor vehicle traffic accident male mortality and population size, England and Wales 1996

Age	**Motor vehicle traffic accidents, number of deaths, males**		**Populations (1996)**		
	England	*Wales*	*England (1000s)*	*Wales (1000s)*	*European Standard*
0 – 14	113	7	4846.2	288.6	22000
15 – 24	589	28	3106.5	184.7	14000
25 – 34	460	32	4050.6	214.0	14000
35 – 44	241	18	3410.1	192.4	14000
45 – 54	192	17	3185.9	192.0	14000
55 – 64	156	13	2362.5	149.3	11000
65 – 74	124	6	1931.0	128.4	7000
75 +	229	11	1236.5	78.8	4000
Totals	2104	132	24129.3	1428.2	100000

Source: Office for National Statistics, Series DH1 No 29, 1996. Mortality Statistics: General, England & Wales. Her Majesty's Stationery Office, London.

D3. (a) Briefly explain the difference between a *relative risk* and an *odds ratio*. Give an example of when each measure is appropriate.

(4)

(b) Wald *et al.* (1986) review the results of four studies of women in the USA examining the relationship of passive smoking to lung cancer.

(i) Calculate the odds ratios of lung cancer associated with passive smoking exposure for each of the four studies separately. Which study shows the highest odds of having lung cancer associated with passive smoking exposure?

(5)

(ii) Use the Mantel-Haenszel procedure to find an estimate and confidence interval for the odds ratio of lung cancer for passive smokers compared to those who are unexposed.

(8)

(iii) What factors need to be borne in mind when drawing inferences from these results?

(3)

Exposure to passive smoking among female lung cancer cases and controls in four studies

Study	**Lung cancer cases**		**Controls**	
	Exposed	*Unexposed*	*Exposed*	*Unexposed*
1	14	8	61	72
2	33	8	164	32
3	13	11	15	10
4	91	43	254	148

Source: Wald, N.J., Nanchalal, K., Thompson, S.G. and Cuckle, H.S. (1986). Does breathing other people's tobacco smoke cause lung cancer? British Medical Journal, ***293****, 1217-1222.*

Turn over

D4. (a) Explain what is meant by the *hazard function* in survival analysis.

The Weibull survival distribution is characterised by the hazard function $h(t) = \lambda\gamma t^{\gamma-1}$. Define and derive the corresponding survival function, *S*(*t*), and the probability density function, *f*(*t*), for the Weibull distribution.

(8)

(b) The table below shows the Kaplan-Meier estimate of the survival function *S*(*t*) for 13 patients for the time to removal of a catheter following a kidney infection. Sometimes the catheter has to be removed for reasons other than infection, giving rise to right-censored observations.

Survival analysis for "Time", the number of days from insertion of catheter until removal

```
Time   Status              Cumulative  Standard  Cumulative    Number
                           Survival     Error      Events    Remaining

  8.0  Catheter removed     .9231       .0739         1           12
 15.0  Catheter removed     .8462       .1001         2           11
 22.0  Catheter removed     .7692       .1169         3           10
 24.0  Catheter removed     .6923       .1280         4            9
 30.0  Catheter removed     .6154       .1349         5            8
 54.0  Censored                                       5            7
119.0  Catheter removed     .5275       .1414         6            6
141.0  Catheter removed     .4396       .1426         7            5
185.0  Catheter removed     .3516       .1385         8            4
292.0  Catheter removed     .2637       .1288         9            3
402.0  Catheter removed     .1758       .1119        10            2
447.0  Catheter removed     .0879       .0836        11            1
536.0  Catheter removed     .0000       .0000        12            0

Number of Cases:  13  Censored:  1  (7.69%)  Events: 12
```

(Source: McGilchrist and Aisbett 1991.)

(i) Use a graphical method based upon the estimated cumulative hazard function for checking whether the data may reasonably be assumed to come from a Weibull distribution.

(8)

(ii) Use the graph to estimate λ and γ, the parameters of the Weibull distribution.

(4)

SECTION E - BIOMETRY

E1. An experiment to compare four seeding rates, 50, 75, 100 and 125 g/unit area, was carried out using two varieties of rice. The field layout consisted of four replicates in randomised blocks, and the yields of saleable grain in kg/ha were as follows (4000 has been subtracted from each field record for ease of computation):

	VARIETY A					VARIETY B					RATE TOTAL
Rate	BLOCK I	II	III	IV	TOTAL	I	II	III	IV	TOTAL	
50	1113	1398	1307	678	4496	1346	1952	719	264	4281	8777
75	1272	1713	1483	749	5217	1424	1861	1307	141	4733	9950
100	1164	831	986	410	3391	1656	1777	1546	585	5564	8955
125	1254	542	919	98	2813	804	848	432	748	2832	5645

Block totals are: I, 10033; II, 10922; III, 8699; IV, 3673.
The grand total of all observations is 33327.
The sum of the squares of all the observations is 42751905.

(i) Carry out an analysis of variance to examine the effects of rates and varieties and their interaction, dividing each of these sources of variation into single-degree-of-freedom components.

[The coefficients of linear, quadratic and cubic components for four equally-spaced levels of a factor are, respectively, $(-3, -1, 1, 3)$, $(1, -1, -1, 1)$ and $(-1, 3, -3, 1)$.]

(11)

(ii) (a) With the aid of a diagram, explain the significant terms in the analysis and comment on any which approach significance.

(6)

(b) The scientist conducting this experiment wishes to follow it by a further experiment designed to fit response curves to discover the optimum seeding rates for each of the two varieties. What rates would you recommend should be used in that experiment? Explain briefly the reasons for your choice.

(3)

E2. A food manufacturer wishes to compare four different recipes (R1 − R4) for making a product, and four different processing times (T1 − T4). Only one complete replicate of the 16 recipe/time combinations can be handled each day, so an experiment is planned to last three days. On each day, the recipes are made up in random order; and, once the supply of the product using one recipe has been prepared, the four times are used in random order before making up the next recipe.

Using an index combining several measurements of the quality of the finished product, the manufacturer wishes to assess which would be the best recipe and time to use in future. A higher value of the index shows higher quality. A summary table of the recipe/time totals for this index is:

Time	1	2	3	4	RECIPE TOTAL
Recipe 1	113.4	156.2	144.0	114.5	528.1
2	168.3	169.6	167.0	165.6	670.5
3	165.4	158.8	173.2	172.4	669.8
4	195.0	195.2	179.7	197.6	767.5
TIME TOTAL	642.1	679.8	663.9	650.1	2635.9

Totals for each day are: Day 1, 965.3; Day 2, 936.8; Day 3, 733.8.
The sum of the squares of the 48 individual observations of this index is 150812.79.

(i) Explain what type of design is being used, and draw a diagram to show how the full experiment might have been arranged over the three days. Why is this design a good one to use for this experiment?

(6)

(ii) Check that the *corrected* total sum of squares is 6063.44, and show that the sum of squares for days is 1991.95.

(1)

(iii) Other corrected sums of squares are:

Recipes, 2429.66;
Recipes + Days + (Recipes × Days), 4972.36;
Times, 68.46;
Recipes + Times + (Recipes × Times), 3006.03.

(a) Complete an analysis of variance and use it to indicate what further studies the manufacturer should carry out.

(7)

(b) Make any further tests that are appropriate, and write a brief report on the results.

(6)

E3. (a) Discuss *two* situations in biometry in which line-transect sampling would be a good method to use.

For each, explain briefly but carefully

(i) what population parameters could be estimated,

(ii) how the sample units should be chosen and the sampling carried out,

(iii) any likely sources of bias that the workers who carry out the sampling would need to take special care to avoid.

(8)

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

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

(4)

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

(4)

(iii) Suppose that this standard error is required to be no greater than 0.1. Also, it costs three times as much to locate each plot as it does to sample a plant (but total cost is less important than precision). Examine how many plots will be required, and how the total costs compare, in each of the cases $r = 10, 5, 4$ and 3, assuming the same variance components as found above. Comment on the results.

(4)

E4. Explain the concept of a *tolerance distribution* in biological assay.

(2)

If a group of animals is being tested for response to a substance injected into the bloodstream, let p_i be the probability that animal i responds to dose d_i.

(i) Show that when the tolerances in the population of animals, from which the group was drawn, are Normally distributed, then

$$p_i = \Phi\left(\beta_0 + \beta_1 d_i\right)$$

where Φ denotes the cumulative distribution function of the standard Normal distribution and β_0 and β_1 are constants which you should express in terms of the parameters of the tolerance distribution function.

[*Note*. This defines the *probit* of p_i as $\beta_0 + \beta_1 d_i$.]

(3)

(ii) An alternative assumption for the tolerance distribution is the *logistic distribution*, whose density function may be written as

$$f\left(u\right) = \frac{\exp\left\{\left(u-\mu\right)/\tau\right\}}{\tau\left[1+\exp\left\{\left(u-\mu\right)/\tau\right\}\right]^2} \;, \qquad -\infty < u < \infty$$

where u denotes the tolerance of an individual, and $\tau > 0$. The mean and variance of this distribution are μ and $\pi^2\tau^2/3$.

Show that, with suitable definitions of β_0 and β_1,

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 d_i\right)}{1+\exp\left(\beta_0 + \beta_1 d_i\right)} \;,$$

and hence logit(p_i) = $\beta_0 + \beta_1 d_i$.

(4)

(iii) Draw a rough sketch showing the relative shapes of the probit and logit transformations, comment on their similarities and differences, and discuss briefly why the logit is often used rather than the probit.

(6)

Question E4 continued on next page

(iv) When might it be valid to use log dose, instead of dose, as d_i ?

Five sets, each of 40 animals, were injected with a serum to protect them against pneumonia. The numbers of deaths were as follows.

Dose	0.0028	0.0056	0.0112	0.0225	0.0450
Number (out of 40)	35	21	9	6	1

Computer output for fitting a model in which the logit of proportion dying was related to $\log_e$(dose) was:

scaled deviance = 2.81;   d.f. = 3;
$\beta_0 = -9.19$ (s.e. 1.255); $\beta_1 = -1.83$ (s.e. 0.255).

Calculate ED50 and ED90 estimates using this model. What exactly do they tell us?

(5)

BLANK PAGE

SECTION F - STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A company dispenses shampoo into bottles with a nominal content of 110ml. The standard deviation of the volume of shampoo dispensed into a bottle when the filling operation is in statistical control is 0.94 ml. Random samples of 4 bottles will be taken from the filling line, at approximately 1 hour intervals, and weighed. The mean weight of the bottles will be subtracted and the estimated weights of shampoo will be converted to volumes by dividing by the density, which is known precisely. The standard deviation of the weights of the containers, expressed as an equivalent volume of shampoo, is 0.25 ml.

(i) The company wishes to set a target volume such that only 1 in 1000 bottles will contain less than 110ml. Suggest a suitable target value stating any assumptions you make.

(3)

(ii) Set up Shewhart mean and range charts for the process, and demonstrate their use with the following data.

Sample number	1	2	3	4
	113.59	111.90	113.08	113.75
volume (ml)	112.16	111.48	113.83	113.95
	112.45	111.69	114.11	118.44
	113.08	113.93	114.78	114.33

[Lower and upper 0.1% and 2.5% control chart factors for the standard deviation, when setting up a range chart for samples of size 4, are

lower: 0.20, 0.59
upper: 3.98, 5.31

respectively.]

(13)

(iii) What will the average run length be if

(a) the process mean is on target, (2)

(b) the process mean is at 111 ml? (2)

F2. A chemical manufacturer splits each delivery of raw material into two batches which are then processed. Four samples of material are randomly selected from the product obtained from each batch. Percentage yields from the last three deliveries are given below.

						mean	*standard deviation*
Delivery 1	*batch 1*	34.3	36.4	33.4	33.4	34.38	1.415
	batch 2	38.7	38.5	43.3	36.7	39.30	2.814
Delivery 2	*batch 1*	33.2	35.2	37.8	35.4	35.40	1.883
	batch 2	35.8	37.1	37.1	39.5	37.38	1.544
Delivery 3	*batch1*	40.0	42.6	39.0	40.7	40.58	1.520
	batch 2	39.2	36.6	36.4	43.3	38.88	3.214

(i) (a) Estimate the within batch, between batches, and between deliveries standard deviations.

(11)

(b) What is the standard deviation of a single sample?

(3)

(c) Bearing in mind your answers to earlier parts, which source of variability most affects the standard deviation found in (b)?

(2)

(ii) You are now told that all samples labelled "batch 1" were processed in Reactor *A* and those labelled "batch 2" were processed in Reactor *B*. Determine a 90% confidence interval for the mean difference in reactor yields. What conclusion do you draw from this interval?

(4)

F3. A manufacturer of rubber gloves is studying the average thickness y of gloves produced by his process. Particular interest is in the effects of latex temperature (factor L) and absolute humidity (factor H). The time for which the process runs (factor T) can also be varied, although the manufacturer doubts that this will have a serious effect on y.

He decides to use two different levels of each factor; L and H are each coded "low" and "high", and "short" and "long" values of T are also included in an experiment. The eight possible combinations of these factor levels are run, once each in random order, and the following data are obtained.

Temperature	*Humidity*	*Time*	*Average Thickness*
Low	Low	Short	183
Low	Low	Long	189
Low	High	Short	200
Low	High	Long	194
High	Low	Short	192
High	Low	Long	194
High	High	Short	197
High	High	Long	193

$\sum y = 1542, \quad \sum y^2 = 297404$.

(i) An analysis of variance table for a main effects model is

Source	DF	SS	MS
Latex Temp (L)	1	12.5	12.5
Abs Humidity (H)	1	84.5	84.5
"Residual"	5	86.5	17.3
Total	7	183.5	

Say briefly how significance tests of the main effects in this model might be justified. What conclusions do you draw from these tests? (4)

(ii) A colleague suggests there is most likely to be an interaction between L and H.

(a) Revise the analysis of variance appropriately. (3)

(b) Draw a diagram to show the mean thickness for the four L/H combinations, and comment on the information it gives. (4)

(c) Because T is not expected to have an effect, no terms involving T are studied. Explain briefly whether this appears to be justifiable. (4)

(iii) Another colleague says that a report will have to contain the result of fitting a suitable response surface model to the data. *Without doing any further calculation*, say what model could be fitted, how it could be tested, and which of its coefficients you would expect to be important. (3)

(iv) The averages quoted for each treatment combination in the experiment are based on a random sample of four items from the batch produced. If one of the objectives of the study is to find operating conditions in which y is relatively insensitive to changes in H, what further analysis might be useful? (2)

F4. Suppose n switching devices are subjected to an accelerated wear regime for T hours. During this period m devices fail and their lifetimes are recorded. Assume lifetimes are independent and have an exponential distribution with mean λ^{-1}.

(i) Derive the maximum likelihood estimator of λ.

(6)

(ii) Twenty switching devices were tested for 168 hours, during which 15 failed. The times to failure were 2, 2, 11, 22, 26, 35, 37, 59, 72, 87, 91, 99, 101, 115, 149.

Estimate the mean lifetime. Also estimate the probability that all the switches failed within 336 hours.

(4)

(iii) Assume the lifetimes of the devices in part (ii) are a random sample from a Weibull distribution.

(a) Estimate the parameters α and β in the Weibull cumulative distribution function

$$F\left(t\right)=1-\exp\left(-\left(t/\beta\right)^{\alpha}\right) \quad \text{for } t \geq 0$$

by a graphical method. (Note: it is acceptable to fit a straight line to your plot by eye.)

You may assume

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

where $t_{(i)}$ is the ith order statistic.

(6)

(b) Estimate the probability that a device fails within 336 hours given that it survived 168 hours.

(4)