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

GRADUATE DIPLOMA IN STATISTICS, 2004

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 title.

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.*

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$.

This examination paper consists of 27 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 of Section A starts on page 2.
There are 24 questions altogether in the paper, 4 in each of the 6 Sections.

SECTION A – STATISTICS FOR ECONOMICS

A1. The following data are collected from *Family Spending 2001 – 2002 (Table 1.1)* in order to estimate an Engel curve relating household expenditure on recreation and culture, y, to total household expenditure, x, both measured in £.

Average weekly household expenditure by gross income decile group, United Kingdom, 2001–2002, £.

	Lowest ten percent	*2nd decile group*	*3rd decile group*	*4th decile group*	*5th decile group*	*6th decile group*	*7th decile group*	*8th decile group*	*9th decile group*	*Highest ten percent*
Expenditure on recreation and culture (y)	15.5	21.8	26.0	39.0	43.4	49.4	63.9	76.6	84.5	119.8
Total expenditure (x)	97.5	99.9	114.2	133.1	139.5	148.9	164.9	184.0	199.0	281.8

For information: $\Sigma y = 539.9$, $\Sigma y^2 = 38679.47$, $\Sigma x = 1562.8$, $\Sigma x^2 = 271935.22$, $\Sigma xy = 100465.78$.

(i) Estimate the linear Engel function $y = \alpha + \beta x$ by fitting

$$y = \alpha + \beta x + u, \qquad \text{where } u \sim \mathrm{N}(0, \sigma^2),$$

by ordinary least squares.

Obtain the coefficient of determination r^2, the estimate s of σ, and the standard errors of the estimators of α and β.

(6)

(ii) Hence estimate the elasticity of y with respect to x, i.e. $\dfrac{dy/y}{dx/x}$ at the mean of x.

(2)

(iii) Test the null hypothesis that $\alpha = 0$. Hence give an economic analysis of your fitted function and your results for parts (i) and (ii) of this question.

(4)

(iv) Test the joint null hypothesis that $\alpha = -33$ and $\beta = 0.6$ against the alternative hypothesis that this null hypothesis is false.

(4)

(v) Estimate $\alpha + 150\beta$, the expected expenditure on recreation and culture given a weekly total expenditure of £150, and calculate a 95% confidence interval for your estimate.

(4)

A2. Statistics of quarterly United Kingdom domestic household final consumption expenditure, in total and on transport in particular, in £million both at constant 1995 prices and at current prices, can be derived from Table 1.7 of *Economic Trends Annual Supplement, 2002 edition*, covering the period 1992 to 2001, i.e. 40 quarters in all. This question is based on calculations made using such data, without seasonal adjustment.

(i) How can the figures be used to compile price index numbers for each of the 40 quarters, for transport expenditure and for total expenditure, and how can the results be used to obtain a series of the price of transport relative to total prices?

(4)

The following notation is used in the remainder of this question.

x_{tpt} denotes quarterly transport expenditure.

x denotes quarterly total expenditure.

p denotes the relative price of transport as found in part (i) of this question.

$t = 0.00, 0.25, 0.50, 0.75, \ldots, 9.75$ for the quarters covered in this analysis.

Q_i, for $i = 2, 3, 4$, equals 1 for the ith quarter and 0 otherwise.

x_{-1} denotes x lagged one quarter. Total expenditure in the fourth quarter of 1991 was used to give x_{-1} for the first quarter of 1992.

The following regressions are fitted in turn by ordinary least squares, with estimated standard errors given in parentheses.

$$\underset{}{x_{tpt}} = \underset{(26195)}{-6835} - \underset{(0.07133)}{0.00488x} + \underset{(265.8)}{214.1p} + \underset{(339.2)}{558.6t} \qquad [a]$$

$s = 1960$, $\bar{R}^2 = 0.453$, $rss = 138{,}231{,}232$, $DW = 2.87$

$$\underset{}{x_{tpt}} = \underset{(7797)}{14953} + \underset{(0.03479)}{0.07701x} - \underset{(66.91)}{87.74p} + \underset{(160.2)}{339.7t} - \underset{(214)}{179Q_2} + \underset{(318)}{2994Q_3} - \underset{(402)}{2249Q_4} \qquad [b]$$

$s = 466$, $\bar{R}^2 = 0.969$, $rss = 7{,}162{,}958$, $DW = 1.82$

$$\underset{}{x_{tpt}} = \underset{(8230)}{16135} + \underset{(0.07002)}{0.10745x} - \underset{(0.07365)}{0.03703x_{-1}} - \underset{(67.83)}{90.00p} + \underset{(168.7)}{363.7t} - \underset{(813)}{573Q_2} + \underset{(1028)}{2503Q_3} - \underset{(905)}{2656Q_4} \qquad [c]$$

$s = 471$, $\bar{R}^2 = 0.968$, $rss = 7{,}106{,}832$, $DW = 1.77$

$$\underset{}{x_{tpt}} = \underset{(3275)}{5653} + \underset{(0.03479)}{0.08354x} + \underset{(153.5)}{273.1t} - \underset{(261)}{200Q_2} + \underset{(306)}{2863Q_3} - \underset{(397)}{2365Q_4} \qquad [d]$$

$s = 470$, $\bar{R}^2 = 0.968$, $rss = 7{,}536{,}197$, $DW = 1.99$

Question A2 is continued on the next page

(ii) Equations [b], [c] and [d] include variables Q_2, Q_3 and Q_4. Why was no variable Q_1 relating to the first quarter of each year used in these equations?

(2)

(iii) By comparing equations [a] and [b], test the null hypothesis that the three quarterly variables Q_2, Q_3 and Q_4 as a set do not affect x_{tpt}.

(2)

(iv) Explain what each of the four regressions shows, and why it is or is not satisfactory as an explanation of quarterly transport expenditure. Which of the four equations do you consider to give the most satisfactory explanation of expenditure on transport? Why?

(8)

(v) If you wished to fit a model similar to your preferred regression but embodying constant elasticity of x_{tpt} with regard to appropriate explanatory variables, what would be its algebraic form? How would you proceed to fit it?

(4)

Turn over

A3. The five sextiles divide a distribution or set of observations into six equal parts, so that, for example, the lowest $16\frac{2}{3}\%$ are less than or equal to the first sextile S_1, and S_3 is the median M.

(i) Why might the sextiles be more useful than the quartiles in analysing income distributions?

(3)

Data in pounds relating to weekly household incomes in the United Kingdom official sample surveys of household expenditure in 1996–1997 and in 2001–2002 are as follows.

1996 – 1997			**2001 – 2002**	
Income	*Number of households*		*Income*	*Number of households*
less than 85	642		less than 115	752 (10.06%)
85 but less than 127	641		115 but less than 175	778 (10.41%)
127 but less than 175	642		175 but less than 246	775 (10.37%)
175 but less than 239	641		246 but less than 327	759 (10.16%)
239 but less than 313	642		327 but less than 425	778 (10.41%)
313 but less than 394	641		425 but less than 527	754 (10.09%)
394 but less than 485	642		527 but less than 648	750 (10.04%)
485 but less than 602	641		648 but less than 809	737 (9.86%)
602 but less than 793	642		809 but less than 1085	699 (9.35%)
793 and over	641		1085 and over	691 (9.25%)
	Total 6415			*Total* 7473

Source: Family Spending 1996–97 and 2001–02, Table 9.7.

(ii) Estimate S_1, M and S_5 for each period.

(3)

(iii) What feature of the distribution is measured by $\dfrac{S_5 - S_1}{M}$ and in what units is the result?

Use your estimates of S_1, M and S_5 to evaluate this for each distribution.

Find unit-free measures of skewness for each distribution.

(4)

(iv) Draw overlapping histograms showing the two data sets above in percentage terms. How have you treated the problem posed by the top, open-ended, class intervals? How else might you have treated it?

(6)

(v) Write a brief account of the data and what can be learnt from your calculations and overlapping histograms.

(4)

Turn over

A4. (i) Consider the second-order moving average process

$$x_t = u_t + 0.6u_{t-1} - 0.3u_{t-2}$$

where $E(u_t) = 0$, $E(u_t^2) = \sigma^2$, $E(u_t u_{t'}) = 0$ for $t \neq t'$.

Find the expectation, variance and autocorrelation function of x_t. (4)

(ii) Now consider the first-order autoregressive process

$$x_t = 0.8x_{t-1} + u_t$$

where $x_0 = 0$ and u_t has the same properties as in part (i) of this question.

Find the expectation, variance and autocorrelation function of x_t. (4)

(iii) Describe possible circumstances in which the processes in (i) and (ii) of this question might arise in economic contexts.

(4)

(iv) Quarterly data relating to the period 1979 to 2001, not seasonally adjusted, of United Kingdom household expenditure on clothing and footwear at 1995 prices were obtained from *Economic Trends Annual Supplement, 2002 edition, Table 1.7*. There were thus 92 observations. The following autocorrelations were found.

lag	*autocorrelation*	*lag*	*autocorrelation*
1	0.707	11	0.418
2	0.696	12	0.626
3	0.646	13	0.366
4	0.875	14	0.355
5	0.594	15	0.311
6	0.581	16	0.511
7	0.533	17	0.262
8	0.749	18	0.253
9	0.479	19	0.213
10	0.466	20	0.405

First differences of the logarithms of the data were taken and the following autocorrelations were found using the resultant 91 values.

lag	*autocorrelation*	*lag*	*autocorrelation*
1	–0.529	11	–0.484
2	0.088	12	0.866
3	–0.531	13	–0.466
4	0.952	14	0.081
5	–0.505	15	–0.459
6	0.086	16	0.821
7	–0.510	17	–0.440
8	0.911	18	0.074
9	–0.487	19	–0.432
10	0.084	20	0.771

What do these results indicate about households' expenditure on clothing and footwear in the period to which the data relate?

(8)

SECTION B – ECONOMETRICS

B1. Suppose that an econometric model is of the form

$$y_i = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \varepsilon_i, \qquad [1]$$

where α_1 and α_2 are known to be 1 and α_3 is known to be 2, and the stochastic terms ε_i are independent and all have the N(0, σ^2) distribution, where σ^2 is constant. Ten observations are made on *y* according to the design matrix

$$\mathbf{X'} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

for which summary statistics are $\sum x_{2i} = 55$, $\sum x_{2i}^2 = 385$ and $\sum x_{2i} x_{3i} = 19$.

(i) An econometrician omits the dummy variable x_3 and fits

$$y_i = \beta_1 + \beta_2 x_{2i} + \varepsilon_i. \qquad [2]$$

Show that his estimate of β_2 is

$$\frac{\sum y_i x_{2i} - 5.5 \sum y_i}{82.5},$$

and hence that its expected value is approximately 1.194.

[You should state clearly any results assumed without proof.]

If the estimated value of σ^2 that he obtains by conventional means is 3.9125, what result should he obtain when testing the null hypothesis that $\beta_2 = 1$?

(9)

(ii) Suppose now that the true model is of the form given in [2] with $\beta_1 = \beta_2 = 1$, i.e. is $y_i = 1 + x_{2i} + \varepsilon_i$, but the econometrician fits the model given in [1], i.e. $y_i = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \varepsilon_i$. Without detailed calculation, explain with reasons how and why you would expect the expectations and variances of the coefficients in his fitted model to differ from those of the true model.

(5)

(iii) A student says that it is better to include a regressor variable which is irrelevant than to risk omitting one which is relevant. Comment briefly on this statement.

(6)

B2. (i) Consider the following distributed-lag model:

$$c_t = 0.5 + 0.3y_t + 0.21y_{t-1} + 0.14y_{t-2} + 0.1y_{t-3} + \varepsilon_t ,$$

where c_t denotes consumption at time t, y_t denotes income at time t and ε_t is a stochastic term with the usual properties.

(a) Suggest why a model of this form might adequately represent the reaction of consumption to changes in income.

(b) Calculate the short-run and long-run multipliers for this model.

(c) What are the main problems in trying to fit finite distributed-lag models?

(10)

(ii) The following (Koyck) model was estimated for import demand d_t and income y_t; estimated standard errors of the parameters are given in parentheses.

$$\begin{array}{rcccccc} d_t & = & 0.1 & + & 0.18y_t & + & 0.79d_{t-1} \\ & & (0.09) & & (0.04) & & (0.27) \\ R^2 & = & 0.55, & & DW & = & 1.98 \end{array}$$

From these results, calculate

(a) the speed of adjustment,

(b) the median lag,

(c) the mean lag,

(d) the long-run multiplier.

(6)

(iii) Briefly review the problems which typically arise in fitting Koyck models.

(4)

B3. Keynes (1936) stated "men are disposed … to increase their consumption as their income increases, but not by as much as the increase in their income". He also argued that the proportion of income saved would increase as income increased, so that the average propensity to consume would decline. Two functional forms are proposed:

$$C_t = \alpha_1 + \alpha_2 Y_t + \varepsilon_t , \qquad [1]$$

$$\log(C_t) = \beta_1 + \beta_2 \log(Y_t) + \varepsilon_t . \qquad [2]$$

(i) Determine for each model the conditions for the marginal propensity to consume (mpc = dC_t/dY_t) to be less than 1 and for the average propensity to consume (apc = C_t/Y_t) to decline when income increases. Show also that under these conditions the mpc is less than the apc.

(7)

Quarterly data are given by Charemza *et al* (*New Directions in Econometric Practice, 2nd edition*) in 1970 prices for UK consumers' non-durable expenditure (C_t) and personal disposable income (Y_t) for the period 1957(1) to 1975(4). In the graph below, Y_t is the upper series and C_t is the lower series, and edited Minitab analysis of the data follows. Dummy variables Q_1, Q_2, Q_3, Q_4 act as indicators of seasonal effects: $Q_i = 1$ for observations relating to the ith quarter and $Q_i = 0$ elsewhere, $i = 1, 2, 3, 4$.

(ii) Check whether the conditions found in part (i) are consistent with the estimated models.

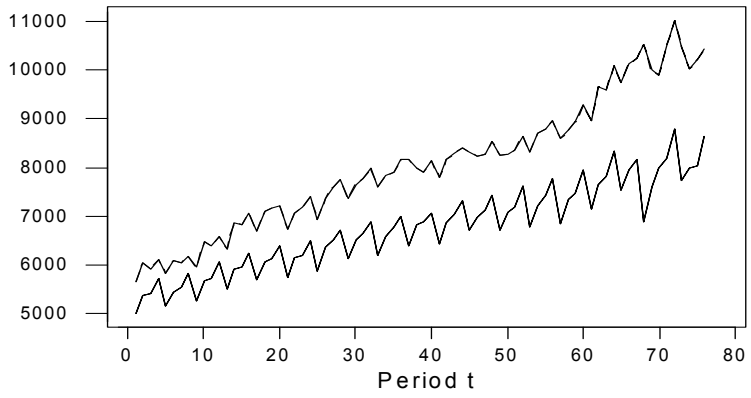
(6)

(iii) Which regression result do you prefer, and why?

(7)

The graph and the Minitab output are on the next page

Graph and Minitab output for question B3




```
MTB > indicators c3 c6-c9   # quarter numbers 1, 2, 3 or 4 are stored in column c3
MTB > name c6 'Q1' c7 'Q2' c8 'Q3' c9 'Q4'
MTB > Regress 'Ct' 4 'Yt' 'Q1' 'Q2' 'Q3';
SUBC>   Constant; SUBC>   DW.

The regression equation is       Ct = 2073 + 0.604 Yt - 434 Q1 - 187 Q2 - 117 Q3
Predictor        Coef     SE Coef          T         P
Constant       2072.8       164.3      12.62     0.000
Yt            0.60443     0.01875      32.24     0.000
Q1            -433.75       72.11      -6.01     0.000
Q2            -187.09       71.59      -2.61     0.011
Q3            -117.04       71.47      -1.64     0.106
S = 219.9       R-Sq = 94.2%     R-Sq(adj) = 93.9%
Durbin-Watson statistic = 1.68

MTB > let c10=log('Ct')          MTB > name c10 'log(Ct)'
MTB > let c11=log('Yt')          MTB > name c11 'log(Yt)'
MTB > name c12 'logfit'     # column c12 contains fitted values for this regression
MTB > Regress 'log(Ct)' 4  'log(Yt)' 'Q1' 'Q2' 'Q3';
SUBC>   Fits 'logfit';    SUBC>   Constant;      SUBC>   DW.

The regression equation is
                log(Ct) = 2.30 + 0.728 log(Yt) - 0.0639 Q1 - 0.0279 Q2 - 0.0169 Q3
Predictor        Coef     SE Coef          T         P
Constant       2.2970      0.1765      13.01     0.000
log(Yt)       0.72829     0.01957      37.21     0.000
Q1          -0.063883    0.009387      -6.81     0.000
Q2          -0.027875    0.009311      -2.99     0.004
Q3          -0.016905    0.009299      -1.82     0.073
S = 0.02861     R-Sq = 95.6%     R-Sq(adj) = 95.4%
Durbin-Watson statistic = 1.74

MTB > let c13=exp('logfit')              MTB > name c13 'fitted'

MTB > let k1=sum(('Ct'-'fitted')*('Ct'-'fitted'))   # sum of squares of (observed - fitted)

MTB > print k1    K1 = 3265342           # values based on this regression but
MTB >                                    # transformed back to Ct-scale
```


Turn over

B4. Write notes on four of the following topics. **(There are 5 marks for each chosen topic.)**

(a) Instrumental variables.

(b) Autocorrelation.

(c) Heteroscedasticity.

(d) Exogenous dummy variables.

(e) Cointegration.

(f) Assessing forecasting models.

SECTION C – OPERATIONAL RESEARCH

C1. (a) A sports car manufacturer has three plants at different sites, all of which manufacture identical cars. The cars are then shipped to four different retailers, each of which can sell 10 cars per month. The production capacity of each plant, and the distances in miles between each plant and each retailer, are shown in the table below.

		Retailer				*Monthly production capacity*
		1	2	3	4	
Plant	1	80	130	40	70	12
	2	110	140	60	100	17
	3	60	120	80	90	11

Each car needs to be shipped individually on a special transporter. The freight cost for each shipment is a fixed cost of £100 plus £5 per mile.

Use the North-West corner method to obtain an initial basic feasible solution of this problem, and then use this to obtain an optimal solution.

(10)

(b) Write down the dual of the following linear programming problem.

Maximize $z = 5x_1 + 4x_2 + 10x_3$

subject to

$$x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0$$
$$2x_1 + x_2 + 5x_3 \leq 40$$
$$4x_1 + x_2 + 3x_3 \leq 30$$

(3)

Use duality theory to determine whether the following are optimal solutions of the primal problem:

(i) $x_1 = 0,\quad x_2 = 15,\quad x_3 = 5$;

(ii) $x_1 = 0,\quad x_2 = 30,\quad x_3 = 0$.

(7)

C2. (i) Consider the $M/G/1$ queue in equilibrium. Write down the Pollacek-Khintchine formula for the average queue length, stating what each symbol in your formula represents.

(4)

(ii) Customers arrive at random at a rate of 25 an hour at an information desk in a railway station. The desk is staffed by a single server. Currently the mean service time is 2.2 minutes and the variance of the service time is 2.1 $(\text{minutes})^2$.

(a) Assuming that the times to deal with separate enquiries are independent, what is the expected queue length, in equilibrium?

(2)

(b) The station managers are not satisfied with this situation and have two choices for improving it. Option A would reduce the mean service time to 1.7 minutes but would increase the variance to 2.5 $(\text{minutes})^2$. Option B would reduce the mean service time to 1.9 minutes but would reduce the variance to 0.1 $(\text{minutes})^2$. Which option would you recommend? What general point does this illustrate about improving the performance of queueing systems?

(6)

(iii) A factory employs one engineer to keep a very large number of machines in running order. The number of machines is so large that it may be treated as an infinite population from which calls for repair occur at random at a rate of 1 per hour. When a machine breaks down, it may require a minor repair (with probability 0.9) or a major repair (with probability 0.1). A minor repair takes half an hour and a major repair takes five hours.

Calculate the mean and variance of the repair time distribution and hence calculate the expected number of machines awaiting repair. How much spare time will the engineer have over a ten-hour period, on average? How do you reconcile this with your answer for the mean queue length?

(8)

C3. (i) Draw a network chart for the project whose activities and prerequisites and personnel requirements are given below. For each activity, find the earliest and latest start time and the earliest and latest completion time. Hence determine the critical path for the project.

(12)

Activity	*Prerequisites*	*Personnel*	*Duration (weeks)*
A	–	3	5
B	–	5	10
C	–	4	1
D	B	1	9
E	D	7	3
F	B	2	8
G	A, F	4	7
H	B	3	10
I	D	9	4
J	C, E	1	5
K	A, F	0	3
L	G, H, I, J	0	8
M	C, E	0	4

(ii) Sketch a Gantt chart for this project. The project manager wishes to minimise the total number of people working on the project at any one time. Assuming that activities cannot be split once started, show that the project can be completed using no more than 14 workers. Indicate a schedule that will achieve this.

(8)

C4. (i) Twenty replications of a simulation experiment were performed to estimate the expected time to service a car. The sample mean time over the 20 replications was 152 minutes, with sample standard deviation 85 minutes. Estimate how many more replications are needed so that the width of the 95% confidence interval for the mean service time is no more than 20 minutes.

(5)

(ii) In the above simulation, the activity of changing the oil was modelled by an exponential distribution with probability density function $f(t) = \mu e^{-\mu t}$, where $1/\mu$ is the mean time taken to change the oil. Using 0.9307 as a random number from a uniform distribution between 0 and 1, obtain an observation from an exponential distribution with mean 5, explaining your reasoning.

(5)

(iii) A village shop is considering opening a small petrol station. There is only space for one pump, and the shop owner wishes to carry out a simulation to see if this will be adequate. Having collected data, the shop owner makes the assessment shown in the following tables.

Time between arrivals (in minutes)	Probability
3	0.05
4	0.20
5	0.35
6	0.25
7	0.10
8	0.05

Filling time (in minutes)	Probability
3	0.10
4	0.20
5	0.40
6	0.20
7	0.10

Use the following random numbers to simulate the operation of this petrol station for one simulated hour, assuming that the petrol station has no customers at the start of the hour:

0.93	0.76	0.27	0.06	0.07
0.47	0.20	0.87	0.83	0.83
0.86	0.07	0.22	0.87	0.49
0.31	0.74	0.63	0.01	0.42

Hence estimate

(a) the mean queueing time per customer,

(b) the mean time spent by a customer in the system.

(10)

Turn over

SECTION D – MEDICAL STATISTICS

D1. (i) Define the sensitivity, specificity, positive predictive value and negative predictive value of a diagnostic test. Comment on why sensitivity and specificity are often preferred to positive and negative predictive values in deciding how good a diagnostic test is.

(6)

(ii) Primary care physicians studying alcohol abuse and dependence need a simple screening test and questionnaire to detect those patients who have drinking problems. The following data come from a study to assess the performance of the CAGE questionnaire in discriminating between medical outpatients who have been confirmed in more extensive tests to be excessive drinkers (alcoholic) or not (non-alcoholic).

CAGE Score	**Number of people and % of total**			
	Alcoholic		*Non-alcoholic*	
0	33	11.2%	428	81.2%
1	45	15.3%	54	10.2%
2	86	29.3%	34	6.5%
3	74	25.2%	10	1.9%
4	56	19.0%	1	0.2%
Total	294	(100%)	527	(100%)

Source: Buchsbaum, D.G., et al (1991). Screening for alcohol abuse using CAGE scores and likelihood ratios. Annals of Internal Medicine vol 115.

(a) Four possible cut-off values for a diagnostic test are CAGE scores of 1 or more, 2 or more, 3 or more, and 4. Estimate the corresponding test sensitivities and specificities.

(4)

(b) Sketch the ROC curve using the four cut-off values from part (ii)(a). Giving your reasons, suggest a suitable cut-off value which gives an appropriate balance between sensitivity and specificity.

(5)

(c) The authors of the report estimated that the overall prevalence of drinking problems among medical outpatients is 36%. Which cut-off gives the best balance of positive and negative predictive values when the prevalence is 36%?

(5)

D2. Thirty patients with chronic osteoarthritis (OA) were entered into a randomised double-blind two-group controlled trial that compared a non-steroidal anti-inflammatory drug (NSAID) with placebo. The trial period was one month. The main outcomes were change (from baseline to one month) in the 10 cm Visual Analogue Scale (VAS) relating to pain, the number of tablets of paracetamol taken during the study and the haemoglobin levels at the end of the study.

(i) What is a double-blind trial? Briefly discuss the advantages and disadvantages of this type of trial.

(4)

(ii) What is a placebo? Discuss whether the use of placebos would be appropriate in such a trial in patients with chronic OA.

(2)

(iii) Briefly describe the different methods of randomisation that might be used in such a trial.

(4)

The mean changes in VAS scores are shown in the table below, together with standard deviations. A positive value implies a reduction (improvement) in pain from baseline to one month.

	Placebo			**NSAID**		
	n	*mean*	*sd*	*n*	*mean*	*sd*
Change in VAS (cm)	15	1.5	2.0	15	3.5	2.5

Source: Campbell, M.J., and Machin, D. (1999). Medical statistics – a commonsense approach. 3rd edition. Wiley.

(iv) Do these data suggest that the NSAID changes the pain scores of OA patients differently from placebo-treated patients one month later? Stating any assumptions you make, perform an appropriate hypothesis test to compare the mean changes in VAS pain scores between the NSAID and placebo treated groups. Comment on the result of this hypothesis test.

(4)

(v) Calculate a 95% confidence interval (CI) for the difference in mean changes in VAS pain scores between the NSAID and placebo groups. Does the CI estimate from these data suggest that patients in the NSAID group have a greater pain change (reduction) at one month than patients in the placebo group? State the assumptions required for this calculation to be valid, and comment on whether the CI gives any additional information to that obtained in part (iv).

(4)

(vi) With reference to VAS pain scores, what is the difference between statistical significance and clinical significance?

(2)

D3. Describe the difference between the direct and indirect methods of age standardising mortality rates. Also discuss when indirect standardisation might be used.

(6)

A census of the City of Southampton and a private census of Southampton University (staff and students) were used to obtain the age distributions of the two populations. The age-specific death rates for England and Wales were also obtained and the results are shown in the table below.

In a one-year period, the observed number of deaths in the City was 2200, and in the University was 6. We wish to compare the mortality of the City and the University.

(i) Calculate the crude death rates for the City and University and comment on the results.

(2)

(ii) Calculate the indirect Standardised Mortality Rate (SMR) for the University.

(3)

(iii) Calculate the indirect SMR for the City.

(3)

(iv) Calculate a 95% confidence interval for this indirect SMR for the City stating any assumptions you make.

(3)

The observed number of deaths in the University was low, so a program Confidence Interval Analysis (CIA), which employs the Poisson distribution, was used for the analysis. This program gave a 95% confidence interval for the indirect SMR for the University (expressed as a percentage) as 18.8 to 112.0.

(v) Discuss the assumptions made in this calculation and the conclusions which may be drawn from it. Does the number of deaths in the University appear particularly low?

(3)

Population of the City of Southampton and the University and national age-specific death rates

Age group	*City*	*University*	*Standard death rates (per 1000)*
0 – 4	25000		0.8
5 – 14	40000		0.4
15 – 24	55000	7500	0.9
25 – 34	50000	1500	1.0
35 – 44	42000	200	2.3
45 – 54	27000	150	7.1
55 – 64	17000	70	20.0
65 – 74	10000	10	52.0
75 – 84	5000		120.0
85+	1000		240.0
All ages	272000	9430	

Source: Campbell, M.J., and Machin, D. (1999). Medical statistics – a commonsense approach. 3rd edition. Wiley.

D4. In an unmatched case-control study of oral contraceptives and breast cancer, the cases were of recently diagnosed and histologically proven breast cancer in women aged 16–50 years in certain hospitals. The controls were married women inpatients in the same hospital, who had certain acute medical or surgical conditions. The control women were interviewed in exactly the same way as the cases. The results of the study are summarised in the table below.

Results of an unmatched case-control study of oral contraceptives and breast cancer

Oral Contraceptives	*Cases*	*Controls*
Ever used	537	554
Never used	639	622
Total	1176	1176

Source: Vessey, M., et al (1983). Oral contraceptives and breast cancer: final report of an epidemiological study. British Journal of Cancer, volume 47.

(i) What is meant by an unmatched case-control study? Discuss the reasons for matching and the advantages and disadvantages of matching relative to unmatched studies.

(8)

We are interested in the relative risk of breast cancer in women taking oral contraceptives compared to women not taking oral contraceptives.

(ii) Explain why we cannot estimate the relative risk directly in a case-control study.

(4)

(iii) What is meant by the *odds* of an event? What is the *odds ratio* for exposure and disease?

(2)

(iv) What is the odds ratio for the occurrence of breast cancer for women using oral contraceptives, relative to those not using oral contraceptives? Comment on the result.

(2)

(v) Calculate an approximate 95% confidence interval for the odds ratio of oral contraceptive use and breast cancer and comment on the results.

(4)

SECTION E – BIOMETRY

E1. (i) Explain the differences between *fixed-effects* and *random-effects* linear models used in the analysis of sets of data. Discuss briefly how the results of an analysis of variance calculation can be used in each case.

(5)

(ii) In a study of techniques for taking a certain measurement in dental research, eight X-ray films are made for each available patient. (Although these are made at different times, the measurement being taken is not expected to change systematically for any particular patient.) Three independent readings of each film are made, under conditions which are controlled as closely as possible.

Substantial patient-to-patient differences are expected, but the main interest is in how accurately the measurements for each patient can be made. The data for one patient are given below (in appropriate units).

Film	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
	8.1	7.2	9.1	9.1	7.1	6.7	10.2	8.1	
	7.6	7.3	9.9	9.6	7.0	6.6	9.9	8.3	
	9.3	7.9	9.6	8.6	6.8	8.4	9.3	9.1	
Total	25.0	22.4	28.6	27.3	20.9	21.7	29.4	25.5	Total 200.8

The sum of the squares of all 24 readings is 1709.82

(a) State an appropriate linear model, and estimate the components of variance between films, σ_F^2, and between readings within a film, σ_R^2.

(7)

(b) The cost of a film is 7 times the cost of taking one reading. Write down a formula for the total cost C of a scheme in which r readings of each of f films are made. Write down also the formula for the variance V of the mean of all the readings obtained using this scheme.

(2)

(c) To minimise either of V or C when the other is given, the product VC may be minimised as a function of r. In the present study, not more than 3 independent readings are possible. By calculating the values of VC for $r = 1$, 2 and 3, using the estimates found in part (a), show that $r = 2$ is the best value to use. Give also a practical reason why this would be preferred to $r = 1$.

If the total cost is to be approximately the same as the cost of the original scheme, show that 9 films could be used. Estimate the value of V if this alternative scheme is used, and compare it with that using the original scheme.

(6)

E2. A plant physiologist is studying the rate of production of leaves by a vegetable crop. Data are available, for the seasons 2000/01 and 2001/02, on the mean number of leaves, y, produced in plots of 10 plants at various times during the season. Different plots were used on each occasion. It is assumed from earlier work that leaf production depends linearly on x, the "accumulated day degrees" since planting, which is a measure of the number of days upon which temperature has risen above a threshold value and the amount by which the threshold was exceeded.

The data are given below, together with extracts from the calculations for fitting separate linear regressions of y on x for the two seasons. [The symbol S denotes *corrected* sums of squares and products, and b_1, b_2 are the regression coefficients for the two seasons.]

2000/01

y_1	3.8	5.3	6.2	6.6	7.2	8.7	10.2	12.0	13.5	15.0	TOTAL 88.5
x_1	4.5	6.0	7.5	8.5	9.5	10.5	13.0	14.5	16.0	18.0	TOTAL 108.0

$\Sigma x_1 y_1 = 1103.85 \qquad \Sigma x_1^2 = 1344.50 \qquad \Sigma y_1^2 = 907.35$

$S_{y_1 y_1} = 124.125 \qquad S_{x_1 y_1} = 148.050 \qquad S_{x_1 x_1} = 178.100 \qquad b_1 = 0.8313$

2001/02

y_2	6.0	6.6	7.8	8.5	9.1	11.0	12.0	12.6	13.3	15.2	TOTAL 102.1
x_2	4.5	5.5	7.0	8.0	9.5	11.0	11.5	13.0	14.0	16.5	TOTAL 100.5

$\Sigma x_2 y_2 = 1132.15 \qquad \Sigma x_2^2 = 1144.25 \qquad \Sigma y_2^2 = 1127.15$

$S_{y_2 y_2} = 84.709 \qquad S_{x_2 y_2} = 106.045 \qquad S_{x_2 x_2} = 134.225 \qquad b_2 = 0.7901$

(i) Plot the data on the same graph, using different symbols for the two seasons.

(4)

(ii) Find the equations of the two regression lines and plot them on the graph.

(4)

(iii) The experimenter wants to combine the data for both seasons, and to examine the hypothesis that they can be explained by fitting a single line to all the data.

(a) Find the equation of this common line, and draw the line on the graph.

(b) The residual sum of squares about this common line is 21.3472. Construct an analysis of variance to examine the hypothesis, and report on your findings. Justify the degrees of freedom for each term in this analysis, and state any assumptions necessary for it.

(9)

(iv) The experimenter now decides to fit two *parallel* lines to the data from the two seasons. You may assume that the least-squares estimate of the common slope of these lines is 0.8136. Write down the equations of these two lines, and *without doing any further calculations* justify briefly the experimenter's choice of this model.

(3)

E3. **Answer any THREE of parts (a) – (e). Each part has equal marks.**

(a) State and explain the circumstances under which a *slope-ratio* biological assay should be preferred to a *parallel-line* assay. Discuss briefly, giving reasons, how a slope-ratio assay should be designed.

(b) Discuss the use of *partial confounding* in designing factorial experiments. Illustrate your discussion by giving examples of schemes for experiments using (i) three and (ii) four 2-level factors in blocks of 4 units, stating the relative precision of estimation of the various main effects and interactions. Why is partial confounding often less useful for larger block sizes?

(c) A response y is calculated as the ratio of two measurements, $y = a/b$. If the mean and variance of a are known to be (μ_1, σ_1^2), and those of b are (μ_2, σ_2^2), and the correlation between a and b is ρ, an *approximate* expression for the variance of y is $\mathrm{Var}(y) = (\sigma_1^2\mu_2^2 + \sigma_2^2\mu_1^2 - 2\rho\mu_1\mu_2\sigma_1\sigma_2)/\mu_2^4$. Under what conditions will this be a satisfactory alternative to using Fieller's Theorem?

Tyre wear has been measured on each of the tyres of n four-wheel-drive agricultural vehicles, of similar type and carrying out similar work. A coefficient of differential wear W has been calculated as $W = (\bar{y}_L - \bar{y}_R)/\{(\bar{x}_L + \bar{x}_R)/2\}$, where $\bar{y}_L, \bar{y}_R$ are the mean wear on left and right front tyres and $\bar{x}_L, \bar{x}_R$ are the mean wear on left and right rear tyres. All measurements are assumed to have the same variance, σ^2, and the amounts of wear on different wheels are assumed to be independent. Construct an expression giving a confidence interval for W. Justify the degrees of freedom for the test statistic used.

(d) Discuss the use of "classical" (polynomial) response surface models for biological work. Explain, with examples, the circumstances when other forms of response surface model would be preferable, and indicate which models might be most suitable for your examples.

(e) In inverse binomial (sequential) random sampling to estimate the proportion p of a certain type of member in a population, a quota of r members of this type is set. The sample items are examined one at a time, and sampling stops as soon as the quota r is met. The number of items sampled to achieve this is n.

Show that the probability mass function of n is

$$P(n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \qquad \text{for } n = r, r+1, r+2, \ldots .$$

Show also that $E[n] = r/p$. Explain why $p^* = r/n$ is <u>not</u> an unbiased estimator of p, and show that $\hat{p} = (r-1)/(n-1)$ <u>is</u> unbiased.

Assuming that an unbiased estimator of $\mathrm{Var}(\hat{p})$ is $p(1 - p)/(n - 2)$, how does the precision of the estimate $\hat{p}$ compare with that from ordinary binomial sampling in which r of special type have been observed in a sample of fixed size n?

State, with reasons, the kind of population and type of member for which this method of inverse sampling is worth considering.

E4. Four laboratories I – IV are carrying out chemical analyses of each of four compounds A – D. The study is spread over eight days (1 – 8) and each laboratory can only deal with one compound each day. The complete study plan is shown in the table below, and you may assume that proper randomisation has been carried out wherever necessary. You may also assume that there are no interactions between any of days, laboratories and compounds.

The response is the departure of the concentration of a particular element from its known theoretical value. These results are shown in the following table (coded into positive whole numbers for ease of analysis).

Table of results of chemical analysis

Day	1	2	3	4	5	6	7	8	Total
Lab I	A: 10	A: 40	C: 4	D: 44	C: 56	B: 6	B: 33	D: 11	204
II	D: 52	C: 73	B: 7	A: 28	A: 44	D: 34	C: 55	B: 2	295
III	B: 45	D: 92	D: 33	C: 56	B: 63	C: 30	A: 44	A: 5	368
IV	C: 53	B: 81	A: 10	B: 51	D: 95	A: 26	D: 88	C: 30	434
Total	160	286	54	179	258	96	220	48	1301

Totals for compounds are A: 207; B: 288; C: 357; D: 449.
The sum of the squares of all 32 observations is 74981.
The sum of the squares of the eight day totals is $160^2 + 286^2 + \ldots + 48^2 = 268837$.

(i) Explain carefully how this design could have been constructed, and write down the linear model upon which the analysis of the results depends, explaining each term in it.

(6)

(ii) Complete an analysis of variance, and examine whether there are differences among laboratories, among compounds, and also between days. Discuss briefly the implications of these results for any future studies of this sort.

(10)

(iii) In this experiment, the compounds were all supplied from the same source. If it had been necessary to obtain them from four different suppliers P, Q, R, S, each of whom was able to supply all four compounds, and possible differences between suppliers also needed to be examined, explain how the design of the study could have been adapted to incorporate this. You may assume that "suppliers" will not interact with any of "days", "laboratories" or "compounds". How (if at all) would the method of construction described in (i) need to be modified?

(4)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. (i) A company produces hanks of a synthetic yarn. Samples of four hanks were taken from the process at hourly intervals over a 20 hour period. The values shown in the table are the weights (in grams) above the lower specification limit. On this scale, the upper specification limit is 150.

Sample					*Mean*
1	69	33	59	36	49.25
2	88	7	61	62	54.5
3	58	7	54	25	36
4	71	38	59	15	45.75
5	61	25	31	27	36
6	70	1	65	31	41.75
7	78	22	42	46	47
8	76	37	26	65	51
9	52	56	54	46	52
10	78	23	80	58	59.75
11	104	8	25	44	45.25
12	68	25	18	32	35.75
13	77	28	33	58	49
14	65	33	62	43	50.75
15	62	0	52	54	42
16	45	31	29	60	41.25
17	64	5	60	77	51.5
18	121	35	70	48	68.5
19	71	35	58	39	50.75
20	27	29	33	39	32
Column total	1405	478	971	905	
Sum of squared values in column	106149	15614	52921	45465	

(a) The average of the 20 within-sample variances is 627.83. Use this to provide a within-samples estimate of σ, the standard deviation of a measured value. When estimating σ, why is it preferable to average the within-sample variances and then take the square root, rather than average the within-sample standard deviations? (3)

(b) Plot the means in a runs chart and show lines at the values $47 \pm k\sigma/\sqrt{n}$ for $k = 1, 2, 3$, where n is the sample size. What do you notice? (6)

(c) Estimate the process performance index. (2)

(ii) Having carried out the analysis in part (i) of this question, you make further enquiries. You are told that the yarn is extruded from a machine with four extrusion heads. Each sample of four is made up of a hank from each head, and columns 2, 3, 4, 5 in the table above correspond to extrusion heads 1, 2, 3, 4 respectively.

(a) Continue with the analysis by calculating column means and variances. (4)

(b) What action would you recommend the company takes? (3)

(c) What process capability index might be achieved? (2)

F2. A production manager in a pharmaceutical chemicals company wishes to maximise the percentage conversion (y%) of a mixture into a medicinal product. The manager has performed an experiment to estimate how four factors, stir rate (A), duration of reaction (B), amount of catalyst (C) and temperature of reaction (D), affect the percentage conversion. A high (+1) and low (–1) value was defined for each factor. The manager used a half replicate of a two-level factorial design, and the results are shown below.

A	B	C	D	y
–1	–1	1	–1	16.1
1	–1	1	1	12.0
1	1	1	–1	17.4
1	–1	–1	–1	10.6
1	1	–1	1	11.8
–1	–1	–1	1	8.6
–1	1	–1	–1	10.4
–1	1	1	1	11.6

The regression model which includes main effects only is estimated as follows.

```
y = 12.3125 + 0.6375 A + 0.4875 B + 1.9625 C - 1.3125 D

The standard error of all the coefficients is 0.6989

Analysis of Variance

Source            DF          SS          MS          F         P
Regression         4      49.745      12.436      3.18     0.184
Residual Error     3      11.724       3.908
Total              7      61.469
```

(i) Suppose we now fit a regression model which has the AB interaction as an additional term.

(a) What is the estimated coefficient of this interaction? (2)

(b) The estimated standard deviation of the errors in the extended model is 0.6755. What is the estimated standard error of each of the coefficients in the extended model? (1)

(c) Complete an ANOVA table for the extended model, including the F-ratio and its statistical significance. (2)

(d) What practical advice would you offer the manager? (2)

(ii) Fit a saturated regression model. Construct a half normal probability plot of the coefficients and comment on the result. (10)

(iii) Suppose the manager claims that the only physically likely interaction is between C and D. What process settings would you advise if the percentage conversion is to be as high as possible subject to all the factors being within the range [–1, 1]? (3)

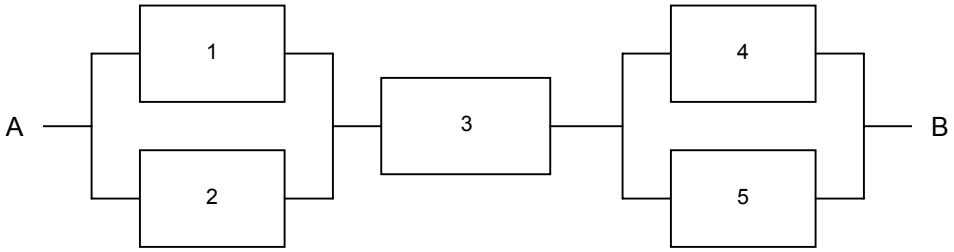
F3. (a) A hearing aid is powered by a single battery of a type which has a mean lifetime of 60 days. The lifetimes of individual batteries are assumed to be independent.

(i) Suppose that the lifetimes of such batteries have an exponential distribution. What is the probability that three batteries will suffice for continuous use for 120 days? (6)

(ii) Suppose instead that lifetimes of such batteries have a Normal distribution with a standard deviation of 20 days. What is the probability that three batteries will suffice for continuous use for 120 days? (4)

[Note. The cumulative distribution function of the gamma distribution with pdf $f(x,\alpha,\lambda)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$, in the case when $\alpha=3$, is $1-\left\{1+\lambda x+\frac{1}{2!}(\lambda x)^2\right\}e^{-\lambda x}$.]

(b) The system shown in the figure below works if there is a path from A to B. For example, the system will work if components 1, 3 and 5 work.



(i) Explain what is meant by the structure function ($\phi(\mathbf{x})$) for a system. Write down the structure function for the system shown in the figure, taking care to define all the symbols used. (3)

(ii) Write down the minimal path sets. (1)

(iii) Write down the minimal cut sets. (1)

(iv) The dual system is defined as the system with structure function

$$\phi_D(\mathbf{x}) = 1 - \phi(\mathbf{1} - \mathbf{x}).$$

Write down the structure function for the dual of the system in the figure, and sketch its block diagram. (4)

(v) Write down the minimal cut sets for the dual system $\phi_D(\mathbf{x})$. (1)

F4. (i) A company has a small electronics workshop, run by one technician, for the repair of equipment on site. Items of equipment fail at a rate of λ per day, and the average time for a repair is $1/\theta$ days. However, the company policy is to send equipment outside for repair if it fails while there is already at least one item awaiting repair in the workshop (in addition to the item under repair). Assume the repair times and times between failures are independent variables from independent exponential distributions.

(a) Write down a suitable state space to represent the numbers of items in the workshop, and draw a diagram to show possible state transitions.

(2)

(b) Explain what is meant by a Markov process.

(2)

(c) For what proportion of time, in terms of λ and θ, would failing items of equipment be sent outside for repair?

(4)

(d) What is the average number of items of equipment sent outside for repair per day, in terms of λ and θ?

(1)

(ii) The company now employs a second technician in the workshop. The technicians do not work together on the same item of equipment. The company maintains the policy of sending equipment outside for repair if there is at least one item of equipment already awaiting repair. Failure rates and repair rates remain the same.

(a) Write down a suitable state space to represent the number of items in the workshop, and draw a diagram to show possible state transitions.

(2)

(b) For what proportion of time, in terms of λ and θ, would failing items of equipment be sent outside for repair?

(5)

(c) Now suppose that λ and θ both equal 3 per day. What is the reduction in the average number of items sent outside for repair per day when the second technician is employed in the workshop? What is the average number of technician hours that are not used for repair work per day?

(4)