EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA IN STATISTICS, 2006

Options Paper

Time Allowed: Three Hours

This paper contains four questions from each of seven 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*
G: *Social, Economic and Financial Statistics**

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 calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

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* ${}^nC_r$.

* Section G of the paper for 2006 will not be released. Candidates who wish to study past examination papers for this Section are advised to refer to 2003 and 2004 rounds of Graduate Diploma papers.

This examination paper consists of 37 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 28 questions altogether in the paper, 4 in each of the 7 Sections.

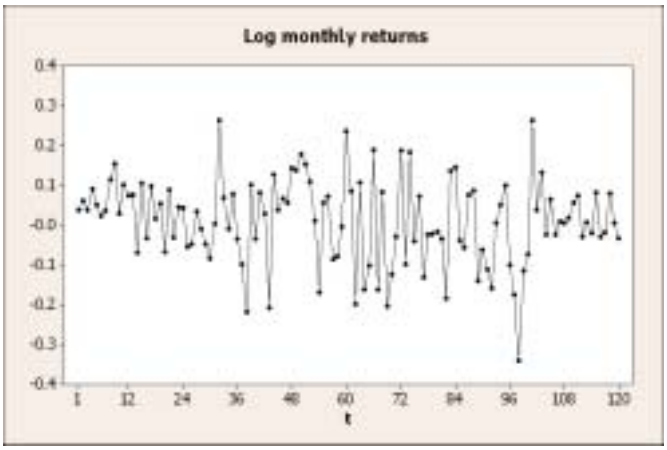
SECTION A – STATISTICS FOR ECONOMICS

A1. Let p_t be the closing price of a share (equity) on the London Stock Exchange on the last trading day of month t, for a UK high-street electrical and audio-visual products retailer. The "log monthly return" on a share is

$$r_t = \log\left(\frac{p_t}{p_{t-1}}\right).$$

Investment analysts seek to advise investors to choose shares where there is an expectation of a positive return.

The data for r_t over the period December 1994 to November 2004 inclusive are plotted below.



(i) The data r_t have mean 0.01072 and standard deviation 0.106570. Assuming that the data are independent, test the hypothesis that the mean log return is zero against the alternative that it is positive.

(3)

(ii) Write down the model for r_t to follow an AR(1) stochastic process, together with the usual assumptions made about the stochastic process.

(4)

(iii) The following AR(1) model with drift is estimated by maximum likelihood with these data.

```
Type           Coef    SE Coef      T       P
AR 1         0.0732     0.0919   0.80   0.427
Constant   0.009928   0.009744   1.02   0.310

Number of observations: 120
SS = 1.34428   MS = 0.01139
```

Explain what each of these numbers means. On this evidence, should investment analysts recommend this company to their clients?

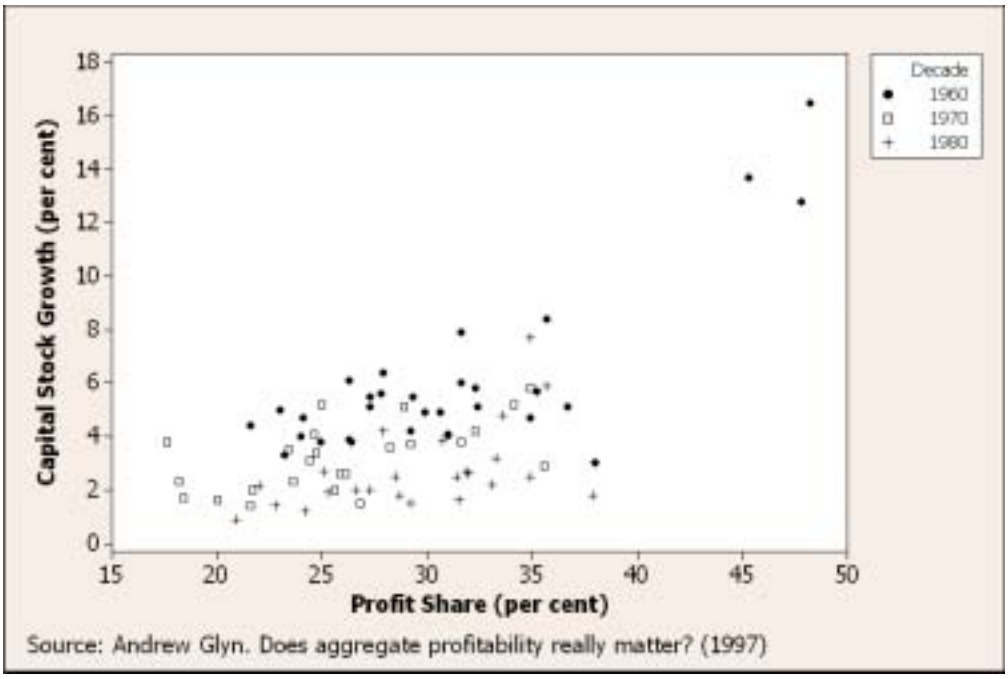
(11)

(iv) What additional statistical analysis of the data would you require to formulate an appropriate model?

(2)

Turn over

A2. In a study on the impact of profitability on capital accumulation, Andrew Glyn[1] gives data on the 5-year growth of manufacturing capital stock (percent per year) and average gross profit shares for some OECD countries, including the UK, for 6 non-overlapping 5-year periods[2]. The data set also includes dummy variables for the decades 1960–69 (`60s`), 1970–79 (`70s`) and 1980–89 (`80s`) and for each country. A plot of these data is given below.



The (edited) output of a statistical package for a multiple regression with the capital stock growth as the dependent variable is shown below. Note that the average gross profit share is denoted by Profsh.

```
Predictor          Coef   SE Coef
Constant        -2.6340    0.9841
Profit Share    0.27635   0.03388
Profsh*70s     -0.18992   0.05051
Profsh*80s    -0.103558  0.009669
70s               3.458     1.394
Japan            3.1302    0.5965
Finland         -1.2121    0.5071

N = 78   S = 1.08288   R-Sq = 84.7%
```

(i) Test each of the coefficients of the fitted regression model for significance. Comment on the explanatory power of the fitted model and on whether this model is satisfactory.

(5)

(Question A2 is continued on the next page)

[1] "Does aggregate profitability really matter", University of Oxford Institute of Economics and Statistics, Discussion Paper 17 (1997).

[2] For some periods, data in some countries are unavailable.

(ii) The profit shares for Japan and the UK in each of three time periods are given in the following table.

Period	1960–64	1970–74	1980–84
Japan	47.8	34.9	35.7
UK	24.9	18.2	22.8

Assuming each period to be representative of its respective decade, use the fitted model to give a point estimate of the growth over five years in manufacturing capital stock for Japan and for the UK in each period.

(7)

(iii) Interpret the estimated coefficients of the dummy variables in this model.

(4)

(iv) What conclusions do you come to about the impact of profitability on capital accumulation?

(4)

A3. Manufacturing corporations in Egypt are classified into 5 sectors:

1. Engineering and Electrical
2. Pharmaceuticals and Chemicals
3. Building materials
4. Spinning, Weaving and Clothes
5. Food

Two researchers[3] sent questionnaires to the finance directors of Egyptian manufacturing corporations. Questionnaires were sent only to manufacturing corporations in sectors 1, 2 and 3. Not all questionnaires were returned.

The researchers stated that only three sectors were selected because "it would be too arduous to survey all manufacturing corporations scattered across different parts of the country".

Manufacturing sectors, Egypt 2001
Numbers of corporations,
numbers of questionnaires distributed and returned

Sector	1	2	3	4	5
Corporations	151	143	111	132	139
Questionnaires					
Distributed	45	39	36		
Returned	35	34	28		

(i) Illustrate, using random numbers from a random number table, how three sectors may be selected from the five sectors with probability proportional to size of sector.

Similarly illustrate how you would select a 30 per cent sample of corporations from each of the chosen sectors.

(8)

(ii) Stating any assumptions made, test whether the proportions of corporations which returned their questionnaires are the same for sectors 1, 2 and 3. State your conclusions clearly, and comment on whether it would be appropriate for the researchers to consider the corporations returning questionnaires to be representative of the sectors.

(6)

(iii) Discuss the implications of selecting only three of the five sectors. State why you feel that the given reason is a valid or an invalid reason for using the sampling method adopted.

(4)

(iv) Describe an alternative sampling procedure which includes corporations from all sectors.

(2)

[3] Ahmed Zakaria Zaki Osemy and Bimal Prodhan, "The Role of Accounting Information Systems in Rationalising Investment Decisions in Manufacturing Companies in Egypt", University of Hull Research Memorandum 30.2001.

A4. Answer four of the following. **(There are 5 marks for each chosen part.)**

(a) What is heteroscedasticity? How may we test for it?

(b) A linear model

$$Y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$$

is fitted to a time series with 35 observations, and a Durbin-Watson *d* statistic of 1.38 is obtained. What evidence is there for serial correlation in the error term, and what other implications might there be for the specification of the model?

(c) For the 17 British general elections from 1945 to 2005, data yield the covariance matrix

	`Final Poll`	`Votes Cast`
`Final Poll`	85.6288	62.9069
`Votes Cast`	62.9069	58.3937

where `Votes Cast` is the amount by which the percentage of votes cast for the winning party exceeds the percentage of votes cast for the second placed party, and `Final Poll` is the amount by which the percentage of those who said they were going to vote for the winning party in the last election poll before election day exceeds the percentage of those who said they were going to vote for the second placed party.

It is argued that in British general elections the last opinion poll of voters taken before each election day is highly correlated with the outcome. Is this so?

(d) The following data for a UK macroeconomic variable are published.

Year	£m at 2000 prices	£m current prices
2000	1047	1047
2005	1085	1101

Compute appropriate quantity and price indices for 2005 for the macroeconomic variable with 2000 as the base year with an index of 100. What kind of indices are they?

(e) Suppose we fit the two models

$$Y_i = \alpha_1 + \alpha_2 x_i + U_i \quad \text{and} \quad \log Y_i = \beta_1 + \beta_2 \log x_i + V_i$$

by ordinary least squares. Discuss how to determine which model to choose on the grounds that it has the better fit.

SECTION B – ECONOMETRICS

B1. Answer four of the following. **(There are 5 marks for each chosen part.)**

(a) Suppose that Y_1, Y_2 and Y_3 are stochastically independent with identical mean β and with corresponding variances σ^2, $4\sigma^2$ and $9\sigma^2$. Show that the ordinary least squares estimator

$$\hat{\beta}_{\text{OLS}} = \tfrac{1}{3}Y_1 + \tfrac{1}{3}Y_2 + \tfrac{1}{3}Y_3$$

and the generalised least squares estimator

$$\hat{\beta}_{\text{GLS}} = \tfrac{36}{49}Y_1 + \tfrac{9}{49}Y_2 + \tfrac{4}{49}Y_3$$

are both unbiased estimators of β, and that $\hat{\beta}_{\text{OLS}}$ is less efficient than $\hat{\beta}_{\text{GLS}}$.

(b) Data for 20 households on consumption expenditure y and income x are ordered by the values of the variable x and then divided into two groups consisting of the first 10 observations and the last 10 observations. The equations fitted by ordinary least squares are

Group 1: $\hat{y}_i = 1.053 + 0.876x_i$, $R^2 = 0.963$, $\hat{\sigma}^2 = 1.210$,

Group 2: $\hat{y}_i = 3.279 + 0.835x_i$, $R^2 = 0.904$, $\hat{\sigma}^2 = 3.154$.

Use the Goldfeld-Quandt test to test for heteroscedasticity, stating any assumptions that you make.

(c) With the information in part (b) above, and the additional information that if the model is fitted to all 20 observations then $\hat{\sigma}^2 = 2.501$, use the Chow statistic to test the claim that high income families have a different consumption function from that of low income families.

(d) Describe the Koyck distributed lag model. Why might an econometrician use such a model and what are the main drawbacks of such a model?

(e) Let $\{X_t\}$ be a time series modelled by

$$X_t = \theta X_{t-1} + U_t$$

where the shocks U_t are independently and identically distributed $\text{N}(0, \sigma^2)$.

Under what conditions is this time series a stationary process? Show that, if $\theta = 1$, X_t can be represented as an infinite sum of all past shocks. Given the experimental values $\{x_t\}$ for T periods, how would you test whether or not it is the case that $\theta = 1$?

B2. (i) Assume the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is non-stochastic and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Given that the OLS estimator of $\boldsymbol{\beta}$ is unbiased, obtain the covariance matrix of the OLS estimator of $\boldsymbol{\beta}$.

(4)

Using 15 years of annual data, planners in San Diego County, USA, estimated by OLS the following model for water consumption:

$$\hat{w} = -326.9 + 0.505\,house + 0.363\,pop - 0.005\,pcy - 17.87\,prw - 1.123\,rain$$
$$\quad (-1.692) \quad (0.910) \qquad (1.378) \qquad (-0.632) \qquad (-1.181) \qquad (-0.825)$$

$$n = 15, \qquad R^2 = 0.93,$$

where

w = total water consumption (million cubic metres),
house = total number of housing units (thousands),
pop = total population (thousands),
pcy = per-capita income (dollars),
prw = price of water (dollars/100 cubic metres),
rain = rainfall in inches

and the values in parentheses are *t* statistics.

(ii) For which slope coefficients in the regression would you expect a positive sign, a negative sign or either sign? Give your reasons.

(3)

(iii) Consistently with your arguments in part (ii), determine the significance of each slope coefficient. Also test the hypothesis that all the slope coefficients are simultaneously zero.

(4)

(iv) Explain your results, drawing attention to any aspects that you consider to be unsatisfactory. Discuss two approaches that might be taken to improve the analysis of water consumption.

(9)

B3. Consider the model

$$Y_t = \beta_0 + \beta_1 M_t + \beta_2 I_t + \beta_3 G_t + \varepsilon_t$$

$$M_t = \gamma_0 + \gamma_1 Y_t + u_t$$

where Y is income, M is money stock, I is investment expenditure and G is government expenditure, and where ε_t and u_t are mutually independent error terms with the usual properties. Annual data for the period 1950–1999 are used in the estimation.

(i) Derive the reduced form equations for this model and comment on the identifiability of the equation for M_t.

(7)

(ii) Suppose you estimate the regression of money stock on income by OLS and obtain

$$\hat{M}_t = 39.0814 + 0.6129 Y_t \qquad R^2 = 0.9957$$
$$(SE = 30.4576) \quad (SE = 0.0090)$$

Comment critically on the properties of the OLS estimates in this estimated equation.

(4)

(iii) Suppose that you decide to estimate the second equation by two-stage least squares. Explain how you would do it.

(4)

(iv) The two-stage least squares regression yields

$$\hat{M}_t = 34.5779 + 0.6144 Y_t \qquad R^2 = 0.9957$$
$$(SE = 30.5416) \quad (SE = 0.0091)$$

Compare these results with those of part (ii). Discuss, with reasons, whether or not the two-stage least squares procedure was worth doing.

(5)

B4. The logit model postulates that the (dependent) random variable Y_i has experimental values $y_i = 0$ or $y_i = 1$ and that

$$P\left(Y_i = 1 \middle| x_i\right) = \frac{e^{z_i}}{1+e^{z_i}},$$

where $z_i = \beta_1 + \beta_2 x_i$, and x_i is an independent variable.

(i) Show that

$$P\left(Y_i = 0 \middle| x_i\right) = \frac{1}{1+e^{z_i}},$$

that the log of the odds ratio is

$$\beta_1 + \beta_2 x_i,$$

and that if $P\left(Y_i = 1 \middle| x_i = 0\right) < P\left(Y_i = 1 \middle| x_i = 1\right)$ then $\beta_2 > 0$.

(5)

A study of 31 contested takeovers in the UK in 1988 postulates that a logit model for the probability of a successful defence, $P\left(Y = 1 \middle| x_1, x_2\right)$, by a target company is

$$P\left(Y_i = 1 \middle| x_{1i}, x_{2i}\right) = \frac{e^{z_i}}{1+e^{z_i}},$$

where $z_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i}$.

The estimated equation is

$$z = \underset{(0.534)}{-0.320} + \underset{(1.440)}{3.460} x_1 - \underset{(1.173)}{2.249} x_2$$

Estimates of the asymptotic standard errors are given beneath the coefficients. The likelihood ratio test yields a χ^2 statistic of 12.083 and the McFadden *R*-square is 0.2854.

The independent variables relate to the documents issued to shareholders by the target company and are defined as

$x_1 = 1$ if there are adverse comments on the financial gearing of the bidder company,
$x_1 = 0$ if there is no such comment,

$x_2 = 1$ if emphasis is placed on photographs or artwork of the principal products or services of the target company,
$x_2 = 0$ if no such emphasis is placed.

(Question B4 is continued on the next page)

(ii) Estimate the probability of a successful defence for each possible combination of values of x_1 and x_2. Given these estimated probabilities, would you advise target companies to include photographs or artwork in their defence documents in the event that they become the target of a hostile takeover?

(7)

(iii) The following table is constructed on the basis that companies are predicted to be successful in their defence if $P\left(Y_i = 1 \middle| x_1, x_2\right) > 0.5$.

Within sample prediction success		
	Actual defence	
Predicted defence	*Successful*	*Failed*
Successful	17	7
Failed	1	6

Having regard to this table, and to the information and results above, appraise critically the fitted model.

(8)

[*Source. The study referred to above is by T.E. Cooke, R.G. Luther and B.R. Pearson, "The information content of defence documents in UK contested takeovers", International Accounting and Finance Research Group, University of Exeter, Discussion Paper No 9009.*]

SECTION C – OPERATIONAL RESEARCH

C1. (a) Staff at a hospital canteen work 12 consecutive hours a day; they begin work at the start of one of the periods listed below. The numbers of staff required to be on duty for each period are as follows.

Period	3am – 9am	9am – 3pm	3pm – 9pm	9pm – 3am
Number of staff required	10	8	9	2

Staff are paid £10 per hour from 6am until 7pm, £15 per hour from 7pm until midnight and £20 per hour from midnight until 6am. In addition, any staff on duty between 6am and 9am are paid an extra £20 for restocking the pantry.

The canteen manager needs to choose staff levels in a way that minimises the total amount paid to staff. Formulate this as a linear programming problem. (Do not solve it.)

(12)

(b) Construct the dual problem associated with each of the following linear programming problems.

(i)
$$\max_{\mathbf{x}} \; -\mathbf{b}'\mathbf{x}$$
subject to
$$\mathbf{A}\mathbf{x} = -\mathbf{c}$$
$$\mathbf{x} \geq \mathbf{0}$$

(ii)
$$\min_{\mathbf{x}} \; \mathbf{c}'\mathbf{x}$$
subject to
$$\mathbf{A}\mathbf{x} \geq \mathbf{b}$$

(iii)
$$\max_{x_1, x_2, x_3} \; 2x_1 + 3x_2 - x_3$$
subject to
$$x_1 + 2x_2 + x_3 \geq 1$$
$$3x_1 - 4x_2 \leq 2$$
$$4x_1 + 2x_2 = 5$$
$$x_2, x_3 \geq 0$$

(8)

C2. Two possible configurations are being considered for a queueing system. The first has a single server who operates at rate μ, while the second also has a single queue but with two servers who each operate at rate $\mu/2$.

In each case there are arrivals at a rate of λ (so the traffic intensity is $\rho = \lambda/\mu$) and a buffer of size 2 (that is up to 2 customers can wait, in addition to those being served).

(i) Give Markov models for each queueing system (transition diagrams will suffice). What assumptions have you made about the inter-arrival and service times?

In each case, write down and then solve the steady state equations.

(12)

(ii) If you were principally concerned with the number of customers lost because the buffer is full, which system would you prefer and why?

(4)

(iii) Suppose $\rho = 1$. If you were principally concerned with the expected waiting time, which system would you prefer and why?

(4)

C3. Consider the function

$$f(x,y) = x^2 - 2xy + \frac{3}{2}y^2 + 4x - y\,.$$

(i) Show that f is convex over $\mathbb{R}^2$.

(5)

(ii) Taking $(x_0, y_0) = (0,0)$ as the initial point, apply one iteration of the steepest descent method to find a new approximation (x_1, y_1) to the minimum of f.

(5)

(iii) Taking $(x_0, y_0) = (0,0)$ as the initial point, apply two iterations of the Newton-Raphson method to find a new approximation (x_2, y_2) to the minimum of f.

(6)

(iv) For a general unconstrained minimization problem, what are the principal disadvantages of the steepest descent and Newton-Raphson methods? Name a better alternative.

(4)

C4. (a) A project consists of activities A, B, …, J, with prerequisites and durations given by the table below.

Activity	*Prerequisites*	*Duration (days)*
A	–	3
B	–	8
C	–	4
D	A	2
E	B	2
F	C	3
G	B, D	6
H	B, D	3
I	E, F	1
J	E, F, H	5

Draw a network representation of these activities. For each event in the network write the earliest and latest event times on the diagram, and hence deduce the critical path.

(12)

(b) A project consists of activities A, B, …, M, with normal and crash durations and the total cost of reduction given by the table below. The costs of reducing for shorter periods are pro rata: if, for example, it costs £20 to reduce a duration by two days, it would cost £10 to reduce it by one day.

Activity	*Normal Duration (days)*	*Crash Duration (days)*	*Cost of Reduction (£)*
A	5	4	25
B	11	8	36
C	7	5	26
D	12	9	12
E	13	10	15
F	6	4	22
G	9	4	30
H	13	7	60
I	7	6	35
J	5	4	10
K	3	2	24
L	12	7	30
M	14	13	8

With normal activity durations, you are given that there are exactly three critical paths:

B – E – L
B – F – J – M
A – D – I – L .

Find how to reduce the duration of the project by one day at minimum total cost.

Suppose that the cost of reducing activity F increases to £30 and the cost of reducing activity L increases to £50. Analyse the effect on your answer.

(8)

SECTION D – MEDICAL STATISTICS

D1. A doctor is designing a randomised controlled trial (RCT) to compare two different treatments of physiotherapy (exercise) for patients suffering from a lung disease called chronic obstructive pulmonary disease (COPD). The two treatments are the standard physiotherapy programme and a new physiotherapy programme with additional exercises. Each patient in the trial is to be randomised to one of these treatments. The doctor wants to know how many COPD patients he will need to recruit.

(i) What type of information would you require from this doctor to estimate a suitable sample size for this proposed RCT?

(6)

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

(2)

One of the ways of assessing the functional ability of patients with COPD is by measuring how far they can walk in six minutes – the six-minute walking test (6MWT). The doctor is proposing to base his analysis on the distances walked, using the 6MWT, at two time-points: baseline, and two months post-baseline.

(iii) Explain briefly what data should be collected, and how they should be analysed. What assumptions would your proposed analysis require, and what would you recommend if these assumptions did not appear to be met?

(8)

(iv) Explain what is meant by an intention-to-treat analysis and a per-protocol analysis. Which one would you recommend for this trial, and why?

(4)

D2. (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.

(9)

(ii) In respiratory medicine, clinicians need a simple diagnostic test to detect those patients with the coalworkers' disease pneumoconiosis. The data in the table show the forced expiratory volume (FEV1), expressed as a percentage of normal values, for a random sample of 40 non-smoking subjects.

FEV1 values (% normal)
for subjects with and without coal-workers' pneumoconiosis

Men with pneumoconiosis $n = 27$								
40	43	47	49	50	50	53	57	58
58	58	62	65	69	71	73	74	75
75	77	78	79	80	87	90	100	105

Men without pneumoconiosis $n = 13$								
60	67	73	75	79	80	83	87	89
100	105	109	115					

Source. A W Musk et al (1981), Br J Indust Med, vol 38.

Four possible cut-off values for a diagnostic test of pneumoconiosis are FEV1 values less than 60% of normal, less than 70%, less than 80% and less than 90%. Estimate the corresponding test sensitivities and specificities.

(6)

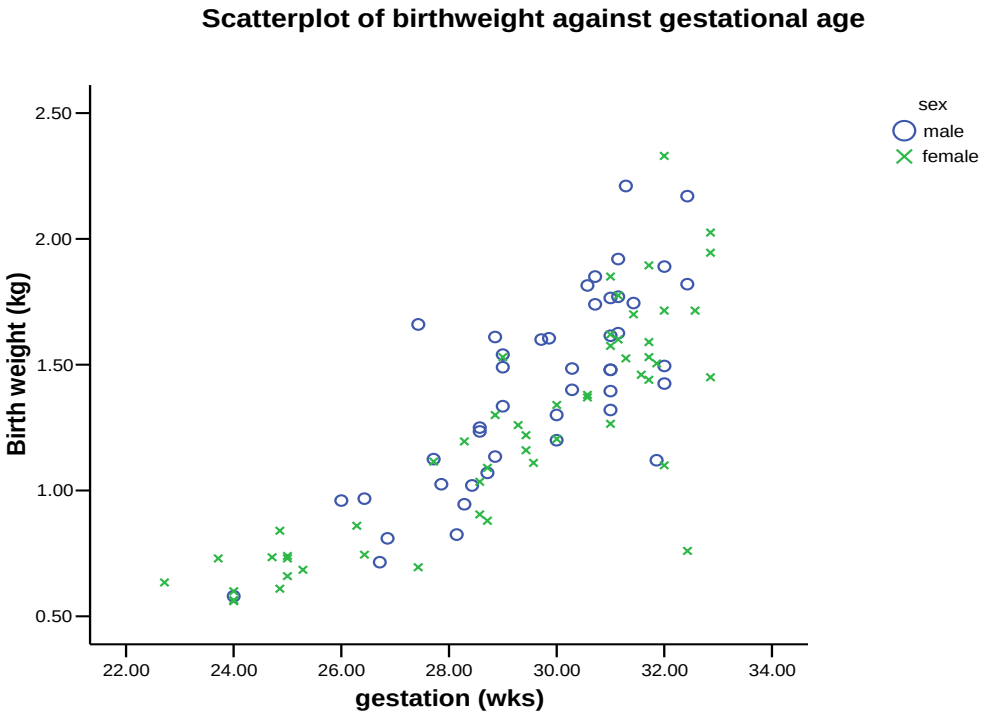
Sketch the ROC curve using the four cut-off values above. Giving your reasons, suggest a suitable cut-off value which gives an appropriate balance between sensitivity and specificity.

(5)

D3. In a study of 98 premature babies, their mother's age (years) was recorded, together with their birthweight (kg), sex and gestation (weeks).

(i) Comment on the scatter plot below, which shows birthweight plotted against gestation, for the male and female babies.

(3)



The tables **on the next page** show abbreviated computer output from two linear regression models, a simple linear regression of birthweight on gestation alone and a multiple linear regression of birthweight on maternal age, sex and gestation. (Note that sex was coded as female = 0 and male = 1.)

(ii) How well can an individual baby's birthweight be predicted from a multiple regression model including maternal age, sex and gestation? Comment on the regression coefficients from this model. (4)

(iii) Compare the results obtained from the multiple regression model with those from the simple linear regression on gestation alone. Which model do you consider more satisfactory? (4)

(iv) Calculate a 95% prediction interval for birthweight from the linear regression on gestation for a baby of 24 weeks gestation. State any assumptions required for the validity of your calculation. (5)

(v) How could we investigate whether it is reasonable to assume that the relationship between birthweight and gestation is linear? (4)

The tables of abbreviated computer output are on the next page

Descriptive Statistics: *birthweight, age, sex, gestation*

	Mean	Std. Deviation	N
Birth weight (kg)	1.3101	.42367	98
Gestation (wks)	29.3353	2.54826	98
Maternal age (yrs)	29.19	6.129	98
Sex	.47	.502	98

Regression 1: *birthweight versus age, sex, gestation*

The regression equation is
birthweight = -2.643 + 0.132 gestation + 0.001 age + 0.116 sex

Predictor	Coef	SE Coef	T	P	Lower bound	Upper bound
Constant	-2.643	.307	-8.601	.000	-3.253	-2.033
Gestation (wks)	.132	.010	13.513	.000	.113	.152
Maternal age (yrs)	.001	.004	.199	.843	-.007	.009
Sex	.116	.050	2.336	.022	.017	.215

S = 0.243 R-Sq = 68.2% R-Sq(adj) = 67.2%

Analysis of Variance

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	11.869	3	3.956	67.105	.000
Residual	5.542	94	.059		
Total	17.411	97			

Regression 2: *birthweight versus gestation*

The regression equation is
birthweight = -2.662 + 0.135 gestation

Predictor	Coef	SE Coef	T	P	Lower bound	Upper bound
Constant	-2.662	.290	-9.179	.000	-3.237	-2.086
Gestation (wks)	.135	.010	13.748	.000	.116	.155

S = 0.247 R-Sq = 66.3% R-Sq(adj) = 66.0%

Analysis of Variance

Predictor	Sum of Squares	df	Mean Square	F	Sig.
Regression	11.546	1	11.546	189.003	.000
Residual	5.865	96	.061		
Total	17.411	97			

Turn over

D4. In a study (*B W Hancock et al (2004), J Clinical Oncology vol 22*) of a new treatment for malignant melanoma (a form of skin cancer), patients were randomised to one of two groups: treatment with low-dose interferon alfa-2a as adjuvant therapy (Intervention), or no further treatment (Control). They were followed up either until the patient died or for up to five years from randomisation. The survival times in years for a random sample of 10 patients from the Intervention group were as follows.

0.91, 1.30, 1.56, 2.59*, 3.74, 3.76*, 4.00, 4.43, 5.0*, 5.0*

(* A star indicates a right-censored observation)

(i) Construct the Kaplan-Meier survival curve for this random sample of 10 patients from the Intervention group and show it on a suitable graph.

(8)

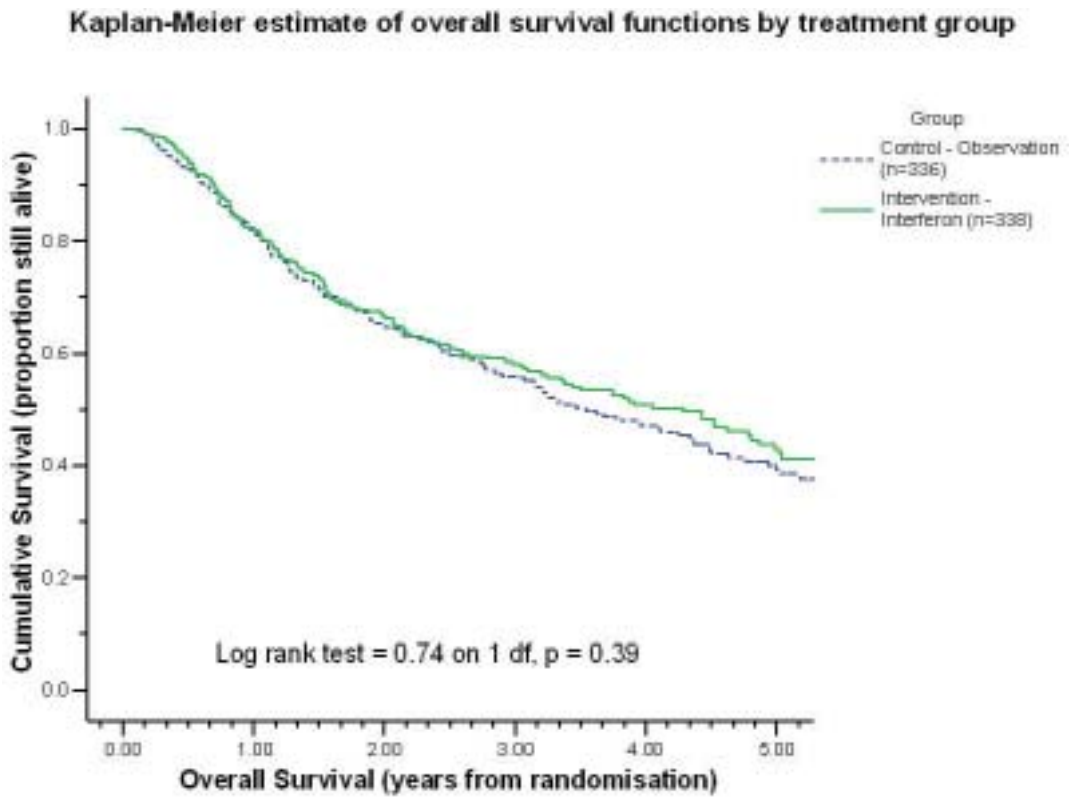
Using Greenwood's formula, calculate the associated standard error for the Kaplan-Meier survival function estimate at 4 years follow-up.

(2)

Calculate an approximate 95% confidence interval for the four-year survival rate for patients on this treatment.

(2)

(ii) The full trial involved 674 patients, with 338 randomised to the Intervention group and 336 to the Control group. The figure below shows Kaplan-Meier estimates of survival functions for the overall survival times for the two treatment groups, and the results of a log-rank test.



(Question D4 is continued on the next page)

Use the diagram to estimate the median overall survival times for the two treatment groups. Is there a difference in the survival of patients in the two treatment groups? Comment on the results of the log-rank test.

(4)

(iii) Previous studies have suggested that age, sex and histology are important factors in predicting overall survival time. Cox proportional hazards regression analysis was used to adjust survival times for these prognostic variables.

The table shows abbreviated computer output from a Cox regression model. The analysis involved a multiple regression of survival time on age (years), sex, histology and group. (Note that sex was coded as male = 0 and female = 1; histology as localised non-metastatic = 0 and metastatic = 1. Similarly, treatment group was coded as Control = 0 and Intervention = 1).

Cox regression – model: age sex histology group

```
No. of subjects =   674            Number of obs    =   674
No. of failures =   351
LR chi2(4)      =    19.59         Prob > chi2      =   0.0006

------------------------------------------------------------------------------
          | Haz.Ratio   Std.Err.       z   P>|z|  [95% Conf. Interval]
----------+-------------------------------------------------------------------
age       | 1.002479   0.0041388    0.60   0.549   0.9944003  1.010624
sex       | 0.7388812  0.0809173   -2.76   0.006   0.5961517  0.9157828
histology | 1.60219    0.2399595    3.15   0.002   1.19462    2.148811
group     | 0.9150073  0.0983535   -0.83   0.409   0.7411898  1.129587
------------------------------------------------------------------------------
```

How well can an individual's overall survival time be predicted from a multiple Cox regression model including age, sex, histology and treatment group? Comment on the regression coefficients from this model.

(4)

SECTION E – BIOMETRY

E1. An experiment was carried out on a particular variety of wheat. Four treatments A – D were used and there were 5 replicates of each, laid out in a randomised (complete) block design. The yields (tonnes/hectare) are given in the following table.

Treatment	*Block I*	*Block II*	*Block III*	*Block IV*	*Block V*	*Total*
A	5.7	6.0	6.7	6.3	7.0	31.7
B	5.8	5.8	6.4	6.4	6.7	31.1
C	6.0	6.5	6.4	6.5	6.7	32.1
D	5.6	5.8	5.1	5.9	6.5	28.9
Total	23.1	24.1	24.6	25.1	26.9	123.8

The sum of the squares of the 20 data values is 770.58.

(i) Analyse the data to investigate whether there is evidence of real differences between the treatments. State carefully the null and alternative hypotheses you are using.

(6)

(ii) The experimenter believes that treatment D may give a mean yield which is different from the overall mean yield of the other treatments. Carry out a suitable further analysis to examine the evidence for or against this belief.

(3)

(iii) Explain the purposes of blocking in field experiments, and comment on whether blocking was useful in this experiment.

(3)

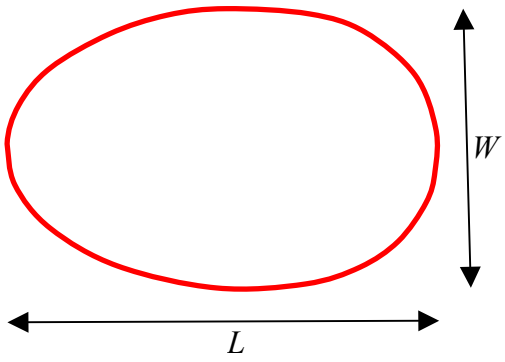
(iv) The experimenter feels that the yield for treatment D in block III is suspiciously low. Discuss briefly any practical circumstances in which it may be valid to treat this item as a "missing value".

(3)

(v) By calculating a suitable residual, or otherwise, assess whether the plot mentioned in (iv) can be regarded as an "outlier". Indicate, without doing any further calculation, how an *exact* analysis of the remaining data, excluding this plot, could be carried out using a covariance method and a dummy variable.

(5)

E2. The diagram shows the cross-section of a hen's egg whose maximum diameter is L (length) and whose diameter of maximum circular section is W (width), where measurements are in centimetres.



It is proposed to find a model to express the volume V (cc) of the egg in terms of its dimensions L and W in logarithmic form where $U = \log_{10}(V)$, $X = \log_{10}(L)$ and $Y = \log_{10}(W)$. The three models proposed are

[1] $U = a + bX + e$

[2] $U = a + bY + e$

[3] $U = a + bX + cY + e$

where e is a Normally distributed residual (error) term having mean 0 and constant variance σ^2.

Twelve eggs were drawn at random from a clutch and their dimensions were as given below.

Egg	*X*	*Y*	*U*
1	0.7659	0.6360	1.765
2	0.7353	0.6198	1.701
3	0.7416	0.6280	1.716
4	0.7600	0.6280	1.737
5	0.7861	0.6239	1.739
6	0.7539	0.6156	1.693
7	0.7747	0.6156	1.725
8	0.7718	0.6239	1.743
9	0.7889	0.6114	1.743
10	0.7659	0.6072	1.710
11	0.7689	0.6156	1.727
12	0.7478	0.6239	1.707

(Question E2 is continued on the next page)

When the three models are fitted by least squares to the data, the parameter estimates (with their standard errors) and the residual sum of squares for each model are given below.

Model	*Constant*	*X*	*Y*	*Residual SS*
1	1.114 (0.236)	0.801 (0.309)		0.002884
2	0.959 (0.447)		1.235 (0.720)	0.003722
3	–0.158 (0.286)	1.024 (0.174)	1.775 (0.357)	0.000769

(i) Inspect the data and comment on any features of interest.

(3)

(ii) Write down the expression which had to be minimised to obtain the parameter estimates shown for Model 1. Given these estimates, state how the residual sum of squares is obtained.

(4)

(iii) Use an appropriate technique to determine which of Models 1 – 3 gives the most suitable representation of the data. Hence estimate the volume of an egg with length 5.8 cm and width 4.1 cm.

(6)

(iv) Using your chosen model from part (iii), calculate the residual which would be obtained for egg 1 after the model had been fitted to the data. Describe briefly methods that could be used to decide whether a model explains the whole set of data well (no further calculation is required).

(4)

(v) When the ellipse whose dimensions correspond to those of the egg in the diagram is revolved about its horizontal axis, the volume of the ellipsoid formed is $(\pi/6)W^2L$. Use this information to comment on the values of the parameter estimates obtained from the three models.

(3)

E3. (i) Explain what is meant by the term *line transect*. Describe circumstances in which line transect surveys might prove useful.

(5)

(ii) The variance of the estimator of a population mean based on a stratified (random) sample is given by the expression

$$V = \sum_{i=1}^{k} \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}.$$

Explain the meaning of the symbols in the expression and explain the conditions under which stratified sampling may be superior to simple random sampling.

(4)

(iii) Explain the meaning of the term *finite population correction factor* and describe the conditions under which its importance is slight.

(3)

(iv) A region contains about 700 farms almost all of which grow rice. A sufficient budget is available to carry out a land survey of about 10% of these farms, and to take crop samples in the field near the time of harvest. Aerial photographs early in the season allowed the region to be classified into areas of good, average and poor soil. Farms and fields are of very irregular shape, although the boundaries of the fields under rice can be determined from the photographs.

The government wishes to estimate

- the total area devoted to growing rice
- the crop yield (in kilograms per unit area).

The crop samples will be taken a week before the main harvest and will be used as a guide to yield.

Describe carefully the steps which should be taken to obtain reliable estimates of these two measurements for the whole region.

(8)

E4. An experiment to test a new insecticide uses k groups of insects. In the ith group there are n_i insects receiving dose x_i, and r_i of them respond to this treatment ($i = 1, 2, \ldots, k$). The theoretical proportion responding at dose i is P_i and the observed proportion is $p_i = r_i/n_i$.

(i) Explain why a logistic model is more useful in practice than an ordinary regression model to describe the proportion responding in each group in this type of experiment. Discuss and compare other suitable alternatives.

(4)

(ii) Show that the logistic model for the proportion P_i responding in each group, which gives the log(odds ratio) as a linear function of x_i, may be transformed to give a relationship of the form

$$P_i = \left(1 + e^{-(\alpha+\beta x_i)}\right)^{-1}.$$

(4)

(iii) Show that the log-likelihood function for this model is of the form

$$l = \text{constant} + \sum_{i=1}^{k} r_i \log P_i + \sum_{i=1}^{k} (n_i - r_i) \log(1 - P_i)$$

and outline how this may be used to estimate the parameters α and β.

(4)

The table shows the numbers (r) of mice out of a total (n) responding to particular concentrations of four different drugs.

Drug	*Dose*	*n*	*r*
1	1.5	103	19
1	3	120	53
1	6	123	83
2	1.5	60	14
2	3	110	54
2	6	100	81
3	0.75	90	31
3	1.5	80	54
3	3	90	80
4	5	60	13
4	7.5	85	27
4	15	90	55
4	10	60	32
4	20	60	44

(Question E4 is continued on the next page)

(iv) A logistic model is fitted to the data, the explanatory variables being the drug (D), the logarithm (**to base 10**) of the dose (L) and their interaction. The resulting residual deviances (scaled deviances) are

Terms	Residual Deviance
(none)	249.96
D	225.70
L	210.50
D, L	4.09
D, L, $D.L$	2.38

Use this information to select the most suitable model and explain your reasoning.

(4)

(v) When the logistic model in (iv) is fitted without interaction, a section of the computer output is as given below.

	Estimate	SE
Constant	-2.268	0.196
Logdose	4.050	0.296
Drug 2	0.392	0.180
Drug 3	2.245	0.223
Drug 4	-1.931	0.223

Note: standard errors are based on a dispersion parameter with the value 1.
Factor parameters for drugs 2, 3 and 4 are differences compared with drug 1 as reference.

Interpret the output. In particular estimate the ED50 (i.e. the value x at which 50% of subjects respond) for drug 1, and compare the effectiveness of the other drugs with that of drug 1.

(4)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. (i) The Laplace distribution consists of suitably scaled back-to-back exponential distributions. It has probability density function (pdf)

$$f(x) = \frac{1}{2}\lambda \exp\left(-\lambda|x-\mu|\right), \qquad \lambda > 0, \quad -\infty < x < \infty .$$

Draw a rough sketch of this pdf and explain why its mean is μ.

Prove that a Laplace distribution whose mean is 0 has variance $2/\lambda^2$. Deduce that the Laplace distribution given above has mean μ and standard deviation $\sqrt{2}/\lambda$.

(4)

(ii) An amplifier is specified to have a gain of 200 units, plus or minus 5%. Suppose that the manufacturing process for these amplifiers leads to a gain with mean 200 units and standard deviation 3 units.

(a) Calculate the process capability index C_p.

(b) If gains have a Normal distribution, what proportion, expressed in ppm (parts per million), of the output will be outside the specification?

(c) What proportion will be outside the specification if gains have a Laplace distribution?

(7)

(iii) (a) Give a general definition of the process performance index C_{pk} and explain how it differs from the process capability index C_p.

(b) An amplifier is specified to have a gain of 200 units, plus or minus 5%. Suppose that the manufacturing process for these amplifiers leads to a gain with mean 204 units and standard deviation 1.7 units. Calculate the process performance index.

(3)

(iv) (a) Let U and V be independent random variables which can take only positive values, and having coefficients of variation C_u and C_v respectively. Define $W = U/V$. Show that, if C_u and C_v are small and C_w is the coefficient of variation of W, then

$$C_w^{\,2} \approx C_u^{\,2} + C_v^{\,2}$$

gives a good approximation for C_w.

(3)

(b) Use this last result to show that the end-points of an approximate 95% confidence interval for C_{pk} calculated from a random sample of n observations from a Normally distributed population are

$$\hat{C}_{pk}\left[1 \pm 1.96/(2n)^{0.5}\right], \text{ assuming that } C_{pk} > 1.$$

[You may assume that the sample standard deviation, calculated from a random sample of n observations from a Normal distribution, is approximately distributed as $N(\sigma, \sigma^2/(2n))$.]

(3)

F2. A company manufactures gas burners for domestic cookers. A machine produces these burners from steel plates. A critical dimension is the length from the base of the burner to the first gas outlet hole. The specification is that this length should be between 67.5 mm and 68.5 mm. When the process was first set up, the standard deviation of length, calculated from extensive records, was 0.12 mm. The machine has been well maintained.

The production manager measures the length on one randomly selected burner from each day's production. Although all the lengths have been within the specification, he notices that the standard deviation of the last 20 lengths is 0.21mm.

(i) Calculate a 90% confidence interval for the standard deviation of length over the past 20 days. List, and comment on, the assumptions you have needed to make. (4)

The manager decides to make a more thorough investigation and, for the next five days, he takes two samples of 4 burners. The first sample is taken near the beginning of the day and the second is taken towards the end of the day. The samples are not four consecutive burners, but are randomly selected from about 15 minutes of production. The data are shown below, together with some summary information.

(ii) Estimate the within sample, between samples within day, and between days standard deviations. (4)

(iii) Set up Shewhart mean and range charts, assuming a target of 68.0, a process standard deviation of 0.12, and a sample size of 4. Plot the 10 means and ranges and comment.

[You may assume that the percentage points of the distribution of relative range, i.e. range/σ, for samples of size 4 are 0.20 (at 0.1%) and 5.31 (at 99.9%).] (5)

(iv) If the mean has moved to 67.9, and the standard deviation remains at 0.12, what is the probability that a sample mean will fall below the lower action line within the next three samples? (3)

(v) Calculate a CUSUM of the means of the samples of size 4, relative to the target of 68.0, and plot the 10 CUSUM values. Set up a V-mask centred on the last point and comment.

[An acceptable V-mask, giving a value of α of about 1 in 440, will have a total width of 10 standard deviations at the point at which it is plotted, and the gradients of the V-mask will be ±5 standard deviations per 10 samples.] (4)

	Day 1 (am)	*Day 1 (pm)*	*Day 2 (am)*	*Day 2 (pm)*	*Day 3 (am)*	*Day 3 (pm)*	*Day 4 (am)*	*Day 4 (pm)*	*Day 5 (am)*	*Day 5 (pm)*
	67.91	68.13	67.90	67.96	67.97	67.87	67.91	68.06	67.72	67.72
	68.01	68.21	68.11	68.00	68.02	67.94	67.96	67.86	67.90	67.81
	68.28	68.05	68.28	68.06	67.97	67.80	68.06	67.93	67.89	68.03
	68.28	68.16	68.03	68.18	68.31	67.86	67.41	67.87	67.88	68.00
Mean	68.12	68.14	68.08	68.05	68.07	67.87	67.84	67.93	67.85	67.89
Variance	0.0358	0.0045	0.0253	0.0092	0.0267	0.0033	0.0842	0.0085	0.0073	0.0223
St dev	0.1892	0.0670	0.1590	0.0959	0.1634	0.0574	0.2901	0.0920	0.0854	0.1494
Range	0.37	0.16	0.38	0.22	0.34	0.14	0.65	0.20	0.18	0.31

F3. An experiment was performed to compare the cut-off times for high voltage switches of three designs (1, 2 and 3) made by two manufacturers (A and B). For each manufacturer, one switch of each design was selected from the stock available in a large warehouse. Each switch was tested twice in a high current and twice in a low current configuration. The order of testing was randomised. The data (cut-off time in hundredths of a second) and an outline ANOVA table follow.

Manufacturer	*Design*	*Low current*		*High current*	
A	1	212	234	283	281
	2	189	190	251	276
	3	221	234	276	272
B	1	204	215	247	251
	2	171	167	240	254
	3	201	190	262	242
	Total	2428		3135	

Totals for designs 1, 2, 3 are 1927, 1738, 1898 respectively.

Totals for current/design combinations are:

	1	2	3
LOW	865	717	846
HIGH	1062	1021	1052

Totals for manufacturer/design combinations are:

	1	2	3
A	1010	906	1003
B	917	832	895

Source of variation	d.f.	Sum of squares	Mean square	*F* value
Manufacturer (M)	1	3151.04		
Current (C)				
Design (D)		2590.08		
M × C interaction		5.00		
M × D interaction				
C × D interaction				
M × C × D interaction	___	243.12		
		27769.46		
Residual (error)	12			
TOTAL	23	28853.96		

(Question F3 is continued on the next page)

(i) Write down a suitable model for the experiment, and explain the meaning and properties of each term in it.

(4)

(ii) Complete the analysis table, including mean squares and F values.

(7)

(iii) Is there evidence of a difference between cut-off times at the two current levels?

(1)

(iv) Test at the 10% level the null hypothesis that there is no difference between manufacturers.

(1)

(v) The mean cut-off times for the switches from manufacturers A and B are 243.25 and 220.33 respectively. Construct a 95% confidence interval for the difference in mean cut-off time between manufacturers A and B.

(3)

(vi) If the $C \times D$ and $M \times C \times D$ interactions were assumed negligible when the experiment was planned, indicate what changes this would make to the analysis and conclusions.

(3)

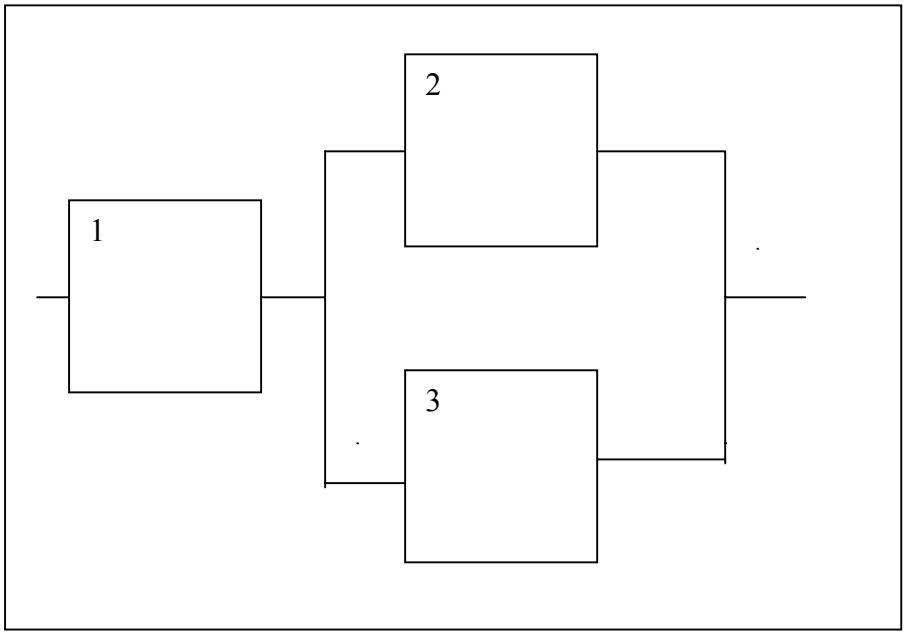
(vii) Comment briefly on whether any conditions mentioned in part (i) seem in doubt on inspecting the data.

(1)

F4. (i) A system is started at time 0, and the probability that it is still working at time t (the reliability function) is $S(t)$. Show that the mean lifetime μ of the system can be written as $\mu = \int_0^\infty S(t)dt$.

(4)

(ii) The diagram below shows a system with three components 1, 2 and 3. The lifetimes of the components have independent exponential distributions whose means, measured in years, are 1, 0.5 and 0.5 respectively. The system works if component 1 and at least one of components 2 and 3 work.



(a) Find the reliability function for this system.

(3)

(b) Use the result in part (i) to calculate the mean lifetime of this system.

(3)

(c) Obtain the hazard function for the system. Calculate the probability that the system fails on the first day (that is, during the first 24 hours) of its second year if it has survived for one year. (Take a year to be 365 days.)

(4)

(d) Deduce that the system behaves like a single component with an exponential lifetime distribution after a long period of time.

(3)

(e) Suppose that the system can be repaired and that repair times have a mean of 300 hours. What is the availability of the system?

(3)

The End