EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA IN STATISTICS, 2005

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* ${}^nC_r$.

This examination paper consists of 35 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. The following equation was estimated by OLS to find the determinants of the natural logarithm of the number of hours per week (h_i) supplied to the labour market by married women, using data from a sample of 560 such women.

Dependent variable: $\log(h_i)$		
Variable	Coefficient	Standard Error
Constant	5.214	
$\log(w_i)$	0.212	0.085
$\log(e_i)$	1.457	0.413
$\log(y_i)$	0.413	0.163
c_i	0.413	0.213
$SSE = 0.248$,	$R^2 = 0.082$	

where the variables are

w_i Hourly wage rate,
e_i Number of years in education,
y_i Husband's income,
c_i Number of children under 5 years of age.

Standard economic theory suggests that, up to a certain point, the number of hours supplied to the labour market responds positively to the hourly wage rate. For married women with children, the number of hours may be reduced by the need to care for the children, but may be increased in order to purchase the services of a child-minder.

(i) At the 5% significance level, test separately the hypotheses that each of the coefficients of $\log(w_i)$ and c_i is zero. (6)

(ii) Test the hypothesis that all slope coefficients are equal to zero. (3)

(iii) Calculate the coefficient of determination adjusted for degrees of freedom, $\bar{R}^2$. If the variable c_i were left out of the equation, would $\bar{R}^2$ increase, stay about the same or decrease? (2)

(iv) Show how dummy variables may be used to define a more general model for the effects of the number of children under 5 years of age on the hours supplied by married women. Why might this be desirable? (5)

(v) Discuss with reasons whether you think this model offers a good explanation of the hours supplied by married women. (4)

A2. A study is to be piloted into educational achievement and employment in towns in the Exeter region (in South-West England). A random sample of 20 persons of working age resident in that region is required, each of whom will be interviewed. The table gives the numbers of persons of working age derived from the 1991 Census of Population for relevant towns containing over 4000 such persons. A list of the names of all persons is available for each town from the electoral register.

There are summer employment opportunities in sea-side towns (Type 1) which are not available in other towns (Type 2).

Persons of working age, 1991		
Town	Type	Persons
Crediton	2	6169
Dawlish	1	10755
Exeter	2	95621
Exmouth	1	28787
Honiton	2	6567
Okehampton	2	4181
Seaton	1	4974
Sidmouth	1	12446
Teignmouth	1	13264
Tiverton	2	15539
Total		198303

(i) Using the table of random digits provided in the statistical tables, draw a simple random sample of 20 persons and identify the towns in which they live. Explain your method clearly.

(10)

(ii) The organiser of the study wishes to use the two types of town as strata, and to obtain a stratified random sample of size 20 allocated in proportion to stratum population. How many sample members should be selected from Type 1 towns and how many from Type 2 towns? Obtain such a sample. In order to save yourself work, you may reuse suitable members of your simple random sample.

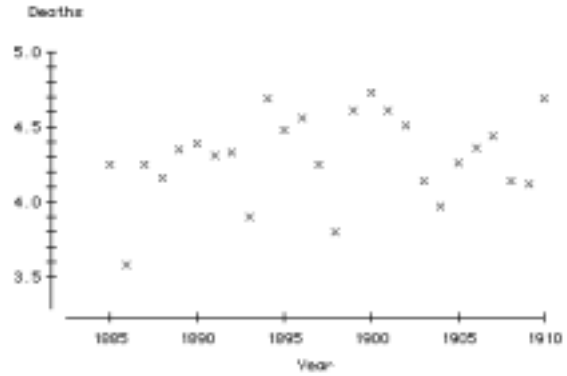
(6)

(iii) Comment briefly on the merits of stratified random sampling as compared to simple random sampling.

(4)

A3. An economic historian examines the hypothesis that the annual numbers of deaths due to industrial accidents, y_t in year t, in the late 19th and early 20th centuries are independent from year to year, showing no systematic pattern over time.

Numbers of persons killed in industrial accidents, UK, 1885-1910, thousands



Source: Department of Employment, British Labour Statistics, Table 200.

Given the evidence in the graph above, she suspects a trend and accordingly differences the data to obtain the series x_t. The autocorrelation and partial autocorrelation functions of the series x_t are as follows.

	ACF	PACF
1	–0.334	-0.334
2	–0.074	-0.208
3	–0.145	-0.288
4	–0.201	-0.498
5	0.396	-0.002
6	–0.077	-0.081
7	0.169	0.187
8	–0.273	-0.082
9	–0.121	-0.102
10	0.024	-0.308

She fits the AR1 model to the x_t using her statistical package and obtains the following (edited) results.

```
Final Estimates of Parameters
Type       Estimate     St. Dev.   t-ratio
AR   1      -0.4307      0.1969      -2.19
Constant    0.02797     0.07114      0.39
Mean        0.01955     0.04972

No. of obs.:  25
Residuals:    SS = 2.90680
              MS = 0.12638  DF = 23
```

(i) Give the statistical model that the economic historian has fitted, and outline briefly how the estimation is carried out.
(5)

(ii) In the light of the results given by the statistical package, is the economic historian correct in her view?
(7)

(iii) The results of fitting a simple linear regression model to the original data are

$$y_t = -11.66 + 0.00842t \qquad R^2 = 0.051 \qquad DW = 1.82 .$$

Do you consider either of these models to be suitable to represent the data above, or do you feel that a different model is needed? Give your reasons. (8)

Turn over

A4. In a contested take-over, soon after the bid for its shares is announced, the directors of the target company issue a defence document to its shareholders to attempt to persuade them not to accept the terms of the bid. Stem and leaf diagrams for the number of pages of such documents for all contested take-overs in the UK for 1988–1990 are given below, by whether the target company failed or succeeded in defeating the take-over bid. The sample means and the unbiased estimators of the variance (using divisor $n - 1$) are also given. (*Source: Cooke T.E., Luther R.G. and Pearson B.R., The Information Content of Defence Documents in UK Hostile Take-over Bids, Journal of Business Finance and Accounting, vol 25(1) and (2), pp 115–143.*)

Defence failed

```
Leaf Unit = 1.0

 0 1246                              Sample mean   27.353
 1 24799
 2 00011244788                       Unbiased estimate of variance   218.175
 3 004467
 4 01248
 5 03
 6 2
```

Defence succeeded

```
Leaf Unit = 1.0

 0 444                               Sample mean   29.511
 1 02245666666667899
 2 0012344578                        Unbiased estimate of variance   563.301
 3 268
 4 000158
 5 346
 6 6
 7
 8
 9
10 5
11
12 4
```

(i) Interpret the stem and leaf diagrams for these data. Use the stem and leaf diagrams to calculate estimates of the median and upper and lower quartiles of each distribution. (9)

(ii) Treating these data as if they were large simple random samples, and using the summary statistics given, test the conjecture that there is no difference in the mean length of defence documents between target companies who are unsuccessful and those who are successful in defeating the take-over bid. (4)

(iii) In the light of the above results, comment critically on the appropriateness of the test procedure you have undertaken in part (ii). (4)

(iv) The length of the defence document is but one aspect of the document. In the light of your economic understanding of contested take-overs, what other aspects of the defence document should be assessed for their relevance to the outcome of a defence? (3)

SECTION B – ECONOMETRICS

B1. (i) Let the proportionate increase r from one period to the next in a time series x_t be defined as $r = x_t/x_{t-1}$. Let the difference operator Δx_t be defined as $\Delta x_t = x_t - x_{t-1}$. Show that $\log(r) = \Delta\log(x_t)$.

(2)

A study of data from the administrative records of the Lancashire Careers Service, UK, by M. J. Andrews, S. Bradley and R. Upward (*Discussion Paper EC10/96, The Management School, Lancaster University*) was used to examine the wages of 845 young people who left school aged 16–18 between 1988 and 1991. About one fifth of school leavers took part in the Youth Training Scheme (YTS) subsidised by the government and about half went straight into work. The study argued that those on a YTS scheme earned less initially than those who went directly to work, but earned more on average after training was finished. Hence the growth in wages of ex-trainees should be higher than that of non-trainees. The authors gave the following regression results.

Dependent variable: $\Delta\log(w_t)$

Independent Variable	Regression (1)	(2)	(3)
D	−0.025 [0.010]	−0.022 [0.020]	−0.021 [0.026]
$\log(w_t{-}1)$	−0.069 [0.000]	−0.066 [0.000]	
$D \times \log(w_t{-}1)$	0.017 [0.203]		
Provider	−0.009 [0.219]	−0.009 [0.210]	0.024 [0.000]
Female	−0.009 [0.134]	−0.009 [0.124]	−0.012 [0.059]
$D \times$ Female	0.020 [0.125]	0.020 [0.113]	0.024 [0.063]
Age17	0.009 [0.417]	0.009 [0.399]	−0.009 [0.425]
Age18	0.053 [0.016]	0.053 [0.017]	0.043 [0.064]
R^2	0.288	0.286	0.222

Notes: Figures in square brackets are two tail p–values.
A constant is included in the regression but not reported.

where w_t is the nominal wage rate at time t and

$D = 1$ if an ex-YTS-trainee, 0 otherwise

Provider = 1 if the employer provided YTS work-experience

Female = 1 if female

Age17 = 1 if aged 17 when taking up first job

Age18 = 1 if aged 18 when taking up first job

[Age16 is the omitted dummy variable].

(Question B1 is continued on the next page)

The authors' first model is

$$\Delta \log\left(w_t\right) = \left(\alpha_0 - 1\right)\log\left(w_{t-1}\right) + \left(\alpha_1 - \alpha_0\right)D \times \log\left(w_{t-1}\right) + \alpha_2 D + \text{other variables},$$

in which imposing the constraint $\alpha_1 = \alpha_0$ gives their second model

$$\Delta \log\left(w_t\right) = \left(\alpha_0 - 1\right)\log\left(w_{t-1}\right) + \alpha_2 D + \text{other variables},$$

and further imposing the constraint $\alpha_0 = 1$ gives their third model

$$\Delta \log\left(w_t\right) = \alpha_2 D + \text{other variables}.$$

(ii) Analyse these regression results by considering which of the restrictions that the authors impose are consistent with the data.

(7)

(iii) For your preferred estimated model, which of the variables, if any, offer a significant explanation of the proportionate increase in the nominal wage rates for the young persons in this set of data?

(8)

(iv) Discuss whether your preferred regression result provides support for the view "that those on a YTS scheme earned less initially than those who went directly to work, but earned more on average after training was finished".

(3)

B2. By using data on real personal consumption expenditure, C_t, and real disposable personal income, Y_t, for annual data over the period 1953–1993, you get the following OLS results.

$$\hat{C}_t = \underset{(\text{S.E.} = 90.99)}{-65.80} + \underset{(\text{S.E.} = 0.009)}{0.916 Y_t}$$

The Durbin-Watson statistic is $d = 0.461$.

(i) Test a null hypothesis of no first order serial correlation in the error term versus an economically plausible alternative hypothesis. State clearly the null and alternative hypotheses. What can you conclude about serial correlation in the error term?

(6)

(ii) What can you infer about the properties of the OLS estimates?

(4)

(iii) Outline how you would use generalised least squares (GLS) to obtain better estimates than those given by OLS in this case. In what sense would the GLS estimates be better?

(6)

(iv) Briefly describe two methods, not based on the Durbin-Watson statistic, of estimating the first-order serial correlation coefficient.

(4)

B3. Suppose you have the following income-consumption model:

$$C_t = \beta_{11} + \beta_{12} Y_t + \varepsilon_{1t}$$
$$I_t = \beta_{21} + \beta_{22} R_t + \varepsilon_{2t}$$
$$Y_t = C_t + I_t + G_t$$

where C is private consumption expenditure, I is private investment, Y is gross national expenditure, R is a weighted average of interest rates and G is government expenditure.

(i) Say, with reasons, which variables are endogenous to the model and which are exogenous to the model.

(3)

(ii) Let the reduced form equations for this model be

$$C_t = \pi_{10} + \pi_{11} R_t + \pi_{12} G_t + \nu_{1t}$$
$$I_t = \pi_{20} + \pi_{21} R_t + \pi_{22} G_t + \nu_{2t}$$
$$Y_t = \pi_{30} + \pi_{31} R_t + \pi_{32} G_t + \nu_{3t} \ .$$

Solve for the coefficients π_{10}, π_{11}, …, π_{32} in terms of β_{11}, β_{12}, β_{21} and β_{22}.

(5)

(iii) Examine the identifiability characteristics of the first and second structural equations and develop indirect least squares (ILS) estimators for β_{11} and β_{12}. Are these the only possible ILS estimators for β_{11} and β_{12}?

(4)

(iv) What would your main concern be if someone attempted to estimate the structural equations by ILS?

(3)

(v) Explain how you would use the two-stage least squares (2SLS) estimator to estimate the first structural equation. Would you prefer this estimator to the ILS estimator? Why or why not?

(5)

B4. Answer four of the following. **(There are 5 marks for each chosen part.)**

(a) Use the following extract of output to discuss the differences between the coefficient of determination, the Schwarz criterion and the Akaike information criterion as measures of fit of a regression.

```
N = 17
K = 3
R-SQUARE = .9513
R-SQUARE ADJUSTED = .9443
SCHWARZ CRITERION = 44.63
AKAIKE INFORMATION CRITERION = 38.53
VARIANCE OF THE ESTIMATE-SIGMA**2 = 30.951
```

(b) Discuss how you would use diagnostic tests for heteroscedasticity in a regression context, and describe a suitable test.

(c) Multicollinearity is mainly an estimation problem; it does not affect the sampling properties of the estimators. Discuss.

(d) Suppose that an estimated regression model omits a variable which is specified in the true model. Suppose further that the experimental correlations between the omitted variable and the included variables in the equation are zero. What is the implication of this for the bias and efficiency of the estimated coefficients? What are the implications if at least one of the correlations is non-zero?

(e) What is a unit root? What are the implications of having a unit root in a time series?

(f) What are distributed lag models? Give an example where it would be advantageous for one to be used and explain why this is so.

SECTION C – OPERATIONAL RESEARCH

C1. A tool firm makes two types of T-square, Type A and Type B. Both products require grinding, calibrating and polishing. Type A products require 4 hours grinding, 1 hour calibrating and 2 hours polishing. Type B products require 2 hours grinding, 2 hours calibrating and 2 hours polishing. The total amounts of time available at the plant for grinding, calibrating and polishing are 52, 36 and 40 hours per week respectively. The firm makes a profit of £12 for each Type A and £15 for each Type B produced.

(i) Formulate the problem of how many T-squares of each type to produce (per week) as a linear programming problem.

(3)

(ii) The final simplex tableau for this problem is given below. Here x_1 and x_2 are the numbers of Type A and Type B T-squares produced; s_1, s_2 and s_3 are the slack variables (in hours) for the grinding, calibrating and polishing constraints; z is the objective.

z	x_1	x_2	s_1	s_2	s_3	RHS
	0	0	1	2	–3	4
	0	1	0	1	–0.5	16
	1	0	0	–1	1	4
1	0	0	0	3	4.5	288

Explain the optimal solution in a manner that would be useful for the tool firm. Include an interpretation of the shadow prices, and explain by how much the total times for each of the three activities could change, while this current solution remains optimal.

(7)

(iii) Using the graph paper provided, draw the feasible region for this problem, identify the vertex corresponding to the optimal solution, and indicate why it is optimal.

(5)

(iv) Suppose that the profit on Type A T-squares increases to £16 each. State what alterations should now be made to the tableau above, and why the current solution is no longer optimal. Use the simplex algorithm to find the new optimal solution.

(5)

Turn over

C2. (i) A gift shop obtains photograph frames of two different sizes from the same supplier. For each size, the price of an individual frame, the cost of placing an order, the holding cost per frame per annum and the annual demand are given in the following table. Demand for both types is steady and shortages must not occur.

Size	*Price (£)*	*Order cost (£)*	*Holding cost (£)*	*Annual demand*
Small	4	10	5	900
Large	15	20	8	320

Orders for small and large frames must be placed separately. The supplier gives a discount of 2.5% on each individual order for which the pre-discount frame cost is at least £1200. Determine the optimal order quantity and the total annual cost for each type of frame.

(8)

(ii) Describe the costs involved in maintaining an inventory.

(2)

What is meant by the term *lead time*?

(2)

What is meant by the term *safety stock*?

(2)

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

(6)

C3. The value of the integral $I = \int_0^1 x^2(1-x)dx$ is to be evaluated by each of three techniques, listed below. In each case, describe the application of the technique to this problem, and estimate how many uniform variates will be needed to ensure that the standard deviation of your estimate of *I* does not exceed 0.001.

(i) Hit-or-Miss Monte-Carlo.

(6)

(ii) Crude Monte-Carlo.

(7)

(iii) Importance Sampling via the pdf $f(x) = 6x(1-x)$. (You may assume that a routine for sampling from this distribution, using one uniform variate, is supplied.)

(7)

C4. Consider a single server queue with arrivals according to a Poisson process of rate λ customers per unit time and independent and identically distributed service times with mean α and variance β.

(i) Define the traffic intensity ρ.

(2)

(ii) Let ξ_n be the number of arrivals while customer *n* is being served. Find $E(\xi_n)$ and hence give a sufficient condition for the queue to reach a stable equilibrium.

(5)

(iii) State the Pollaczek-Khintchine formula for the expected number of customers in the queue in equilibrium.

(3)

(iv) What is the expected time spent in the queue in equilibrium? State clearly any results you appeal to.

(4)

(v) Suppose that $\lambda = 2$, that the cost of providing the server is £2/α per minute (whether or not the server is busy), and that the cost of keeping a customer waiting before service is £1 per minute. Write down the overall mean cost of this system in equilibrium. Hence deduce, near $\alpha = 1/3$, for what values of β it is worthwhile increasing the value of α.

(6)

Turn over

SECTION D – MEDICAL STATISTICS

D1. (i) In a study to compare two treatments for venous leg ulcers, patients were randomised to receive one of two treatments: treatment at a dedicated community leg-ulcer clinic (Intervention) by specially trained nurses, or normal home treatment (Control) by district nurses. They were followed up until either the initial leg ulcer healed or one year from randomisation. Healing times (weeks) for a random sample from the Intervention group were as follows.

3 5 8 10 11* 27 39 52* 52* 52*
(* A star indicates a right-censored observation.)

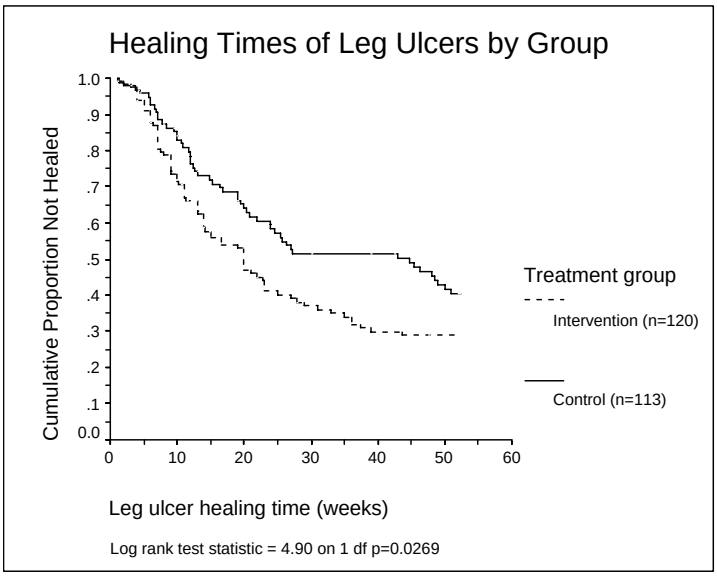
Explain what is meant by a *right-censored* observation. Also explain the meaning of the term *hazard function*.

(4)

(ii) Construct the Kaplan-Meier survival curve for the Intervention group and show it on a suitable graph.

(6)

(iii) The full trial involved 233 patients, with 120 randomised to the Intervention group and 113 to the Control group. The figure below shows Kaplan-Meier estimates of survival functions for healing times for the two treatment groups, and the results of a log-rank test; the test statistic was 4.90 on 1 d.f. ($p = 0.0269$).



Source: Morrell, C.J., et al (1998). Cost-effectiveness of community leg ulcer clinics: randomised controlled trial. British Medical Journal vol 316.

Use the diagram to estimate the median leg ulcer healing times for the two treatment groups. Is there a difference between the survival patterns of the two treatment groups? Comment on the results of the log-rank test.

(4)

(Question D1 is continued on the next page)

(iv) Previous studies have suggested that age, initial ulcer size, ulcer duration and history of deep vein involvement are important factors in predicting leg ulcer healing time. Cox proportional hazards regression analysis was used to adjust healing times for these prognostic variables.

The following tables show abbreviated computer output from two Cox regression models. In Analysis 1, a simple regression of ulcer healing time on treatment group alone was performed. Analysis 2 involved a multiple regression of ulcer healing time on age (years), ulcer duration (months), base area (cm^2), deep vein involvement (DVT) and treatment group. (Note that DVT was coded as "No deep vein involvement" = 0 and "Deep vein involvement" = 1. Similarly, treatment group was coded as 0 = Control and 1 = Intervention.)

Analysis 1: Cox regression – model: group

```
No. of subjects =    213            Number of obs   =       213
No. of failures =    129
LR chi2(1)      =   4.85            Prob > chi2     =    0.0276
```

	Haz.Ratio	z	P>|z|	[95% Conf.	Interval]
group	1.480299	2.18	0.029	1.041186	2.104605

Analysis 2: Cox regression – model: age basearea duration dvt group

```
No. of subjects =    213              Number of obs   =       213
No. of failures =    129
LR chi2(5)      =  48.25              Prob > chi2     =    0.0000
```

	Haz.Ratio	z	P>|z|	[95% Conf.	Interval]
age	1.006661	0.82	0.413	.9907812	1.022794
basearea	.9642388	-4.04	0.000	.9473715	.9814064
duration	.9967765	-1.39	0.165	.9922462	1.001327
dvt	.9172132	-0.39	0.695	.5958852	1.411816
group	1.651079	2.75	0.006	1.15515	2.35992

(a) How well can an individual's leg ulcer healing time be predicted from a multiple Cox regression model including age, base area, DVT, ulcer duration and treatment group? Comment on the regression coefficients from this model.

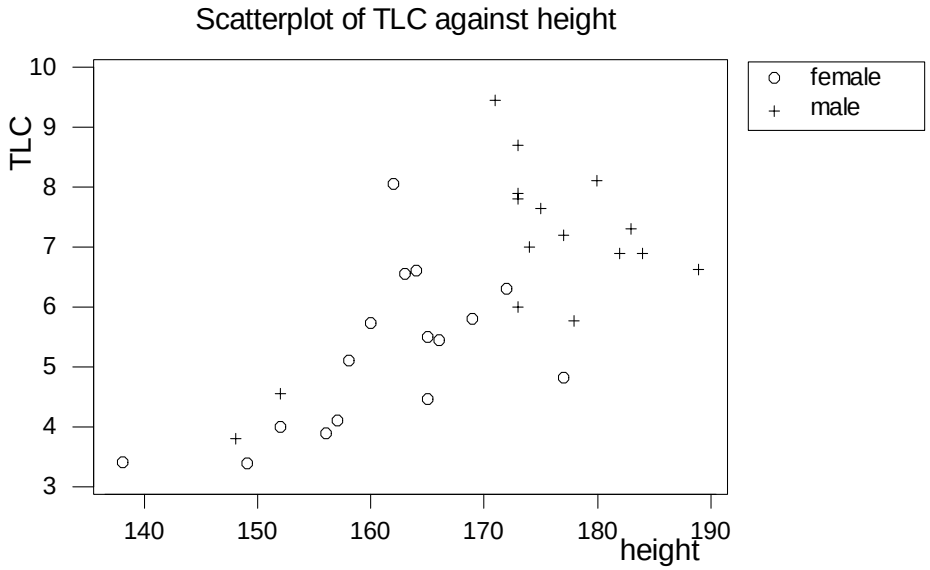
(4)

(b) Compare the results obtained from the multiple regression model with those derived from the regression on treatment group alone.

(2)

D2. For lung transplantation, it is desirable for the donor's lungs to be of a similar size to those of the recipient. Total Lung Capacity (TLC) is difficult to measure, so it is useful to be able to predict TLC from other information. In a study, pre-transplant TLC, age (years), sex and height (cm) of 32 recipients of heart-lung transplants were recorded.

(i) Comment on the scatterplot below, which shows TLC plotted against height for the two sexes. (3)



Adapted from: Otuluna, B., et al (1989). The effect of recipient lung size on lung physiology after heart lung transplantation. Transplantation vol 48.

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

(ii) What is the correlation between TLC and height? (1)

(iii) How well can an individual's total lung capacity be predicted from a multiple regression model including age, sex and height? Comment on the regression coefficients from this model. (4)

(iv) Compare the results obtained from the multiple regression model with those derived from linear regression on height alone. (4)

(v) Calculate the 95% prediction interval from the linear regression on height for someone of height 165 cm. (4)

(vi) How could we investigate whether the relationship between lung capacity and height is the same for males and females? (4)

The tables are on the next page

Descriptive Statistics: *TLC, age, sex, height*

Variable	N	Mean	Median	StDev
TLC	32	6.088	6.150	1.624
age	32	28.41	28.50	10.52
sex	32	0.5000	0.5000	0.5080
height	32	167.44	170.00	11.95

Regression 1: *TLC versus age, sex, height*

The regression equation is
TLC = - 8.54 - 0.0250 age + 0.697 sex + 0.0895 height

Predictor	Coef	SE Coef	T	P
Constant	-8.544	3.679	-2.32	0.028
age	-0.02502	0.02353	-1.06	0.297
sex	0.6970	0.4994	1.40	0.174
height	0.08955	0.02455	3.65	0.001

S = 1.156 R-Sq = 54.2% R-Sq(adj) = 49.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	44.305	14.768	11.05	0.000
Residual Error	28	37.407	1.336		
Total	31	81.712			

Regression 2: *TLC versus height*

The regression equation is
TLC = - 9.74 + 0.0945 height

Predictor	Coef	SE Coef	T	P
Constant	-9.740	2.991	-3.26	0.003
height	0.09453	0.01782	5.30	0.000

S = 1.186 R-Sq = 48.4% R-Sq(adj) = 46.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	39.549	39.549	28.14	0.000
Residual Error	30	42.163	1.405		
Total	31	81.712			

D3. Below are the results of a randomised double-blind placebo-controlled clinical trial examining whether patients with chronic fatigue syndrome (CFS) improved six weeks after treatment with intramuscular magnesium. The group who received magnesium were compared with a group who received a placebo. The outcome recorded was whether the patient felt better.

Treatment	*Felt better*	*Did not feel better*	*Total*
Magnesium	12	3	15
Placebo	3	14	17

(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 the use of placebos in general. Would the use of a placebo be appropriate in such a trial in patients with CFS?

(4)

(iii) Do the data in the table suggest that the patients receiving magnesium had a better outcome than the placebo-treated patients after six weeks? Stating any assumptions that you make, perform an appropriate hypothesis test to compare the proportions feeling better in the magnesium- and placebo-treated groups. Comment on the results of this hypothesis test and the assumptions necessary for it.

(6)

(iv) Calculate an approximate 95% confidence interval (CI) for the difference in the proportions feeling better in the magnesium- and placebo-treated groups. Does the CI estimated from these data suggest that patients in the magnesium group have a better outcome at six weeks 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 (iii).

(6)

D4. The data in the table below describe an unmatched case-control study of lung cancer as related to tobacco consumption. The 483 cases and 447 controls were cross-classified according to smoking status and sex, with the following results.

Results of an unmatched case-control study of smoking status and lung cancer

MALES	**Smoking status**		
	Smokers	*Non-smokers*	*Total*
Cases	58	245	303
Controls	6	271	277
Total	64	516	

FEMALES	**Smoking status**		
	Smokers	*Non-smokers*	*Total*
Cases	31	149	180
Controls	7	163	170
Total	38	312	

We are interested in estimating the relative risk of lung cancer in smokers compared with non-smokers.

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

(4)

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

(2)

(iii) Ignoring sex, what is the odds ratio for the occurrence of lung cancer for smokers, relative to non-smokers? Comment on the result.

(2)

(iv) Calculate the Mantel-Haenszel estimate of the odds ratio for occurrence of lung cancer for smokers relative to non-smokers, allowing for sex. Calculate a 95% confidence interval for this odds ratio.

(8)

(v) Perform a test of the null hypothesis that smoking status is unrelated to occurrence of lung cancer. Comment on the results of all your calculations.

(4)

SECTION E – BIOMETRY

E1. A half-replicate of a 2^5 factorial design was used to investigate five factors A – E, which were fertiliser treatments and crop protection compounds. Each factor was used at two levels, low and high. Earlier studies suggested that A and C were unlikely to interact, and therefore the experimenter assumed that interactions involving AC (i.e. ABC, ACD and ACE) would also be negligible. The half-replicate was constructed by equating ABC with DE. Data from the experiment, in suitably coded units, were as follows.

Treatment combination	*(1)*	*ab*	*ac*	*bc*	*ad*	*bd*	*cd*	*abcd*
Response	61	61	56	54	61	94	66	98
Treatment combination	*ae*	*be*	*ce*	*abce*	*de*	*abde*	*acde*	*bcde*
Response	63	70	59	65	44	77	42	81

The sum of these observations is 1052, and the sum of their squares is 72776.

(i) Write down the defining contrast (or congruence) for this scheme, and use it to list all the alias sets. Indicate which main effects and interactions can be estimated from these alias sets, and state clearly what assumptions must be made when carrying out the estimation. (5)

(ii) A partial Analysis of Variance table is as follows. Copy the table and complete it as far as possible. Show how you compute the missing entries. (4)

Source of Variation	*Degrees of Freedom*	*Sum of Squares*
A		2.25
B		1369.00
C		6.25
D		
E		156.25
AB		6.25
AC		4.00
AD		4.00
AE		4.00
BC		0.25
BD		992.25
BE		
CD		64.00
CE		1.00
DE		625.00
Residual		
TOTAL		

(iii) Explain whether or not there are any alias sets which can provide degrees of freedom for residual ("error").
(2)

(iv) The experimenter now argues that since A and C were not expected to interact, and also the analysis has shown they are not themselves significant either, all the remaining two-factor interactions involving either A or C can be used to form the residual. List the alias pairs which would be used as residual if this argument were to be followed. (5)

(v) Comment on the experimenter's argument, and modify it if necessary to propose a method which allows you to complete a reasonable analysis. Explain carefully how the CD interaction should be studied, and outline *(without any further calculation)* what else should be included in a full report on the results. (4)

E2. **Answer any THREE of parts (a) – (e). Each part has equal marks.**

(a) Define a *linear contrast* among the treatment totals in an analysis of variance of data from an orthogonal design in which all treatments are replicated r times and explain, with examples, the circumstances under which

(i) a complete set of orthogonal contrasts will be useful,

(ii) the contrasts of greatest practical interest will not be orthogonal.

Computer packages often produce, as standard output, the result of carrying out a multiple comparisons procedure on the treatment means. Discuss briefly whether there are situations where this is not useful or appropriate.

(b) Outline the use of discriminant analysis in biometry, illustrating your answer by describing a practical example.

(c) A scientist has collected data from an experiment laid out in randomised complete blocks, sampling each plot r times. He proposes to analyse the data using the linear model

$$y_{iju} = \mu + \tau_i + \beta_j + \varepsilon_{iju}$$

where y_{iju} is the response from a sample on a plot, i ($i = 1, 2, \ldots, v$) denotes the treatment applied to the plot, j ($j = 1, 2, \ldots, b$) denotes the block containing the plot, and u ($u = 1, 2, \ldots, r$) identifies each individual sample. The treatment effects $\{\tau_i\}$ and block effects $\{\beta_j\}$ are fixed effects, and all $\{\varepsilon_{iju}\}$ are independent, identically Normally distributed with mean 0 and variance σ^2. Discuss critically the assumptions required for this analysis to be valid, and explain possible alternative analyses if you consider it unwise to make one or more of these assumptions.

(d) Explain, with examples, the use of nonlinear curves and response surfaces in biometry.

(e) Suppose that your department has been asked to organise a land use survey in a large region of a developing country. Write short notes on what use might be made of land maps, aerial photographs and climatic records (e.g. temperature, rainfall) in planning the survey and deciding the methods of data collection. Under what circumstances might the ratio and regression methods of analysis be useful in estimating totals and proportions?

E3. An experiment is to be designed to test the effects of four levels of nitrogen fertiliser and five growth regulation treatments on the yield of rapeseed. The field to be used for the experiment is known to be of variable fertility, and can conveniently be divided into 60 experimental units. The effect of nitrogen has been examined previously on an adjacent site, but its interaction with the growth regulation treatments is of particular interest in this experiment.

(i) Comment on any advantages a split-plot layout will have over a completely randomised factorial design for carrying out this experiment. Describe how a suitable split-plot layout may be constructed and arranged for this experiment, and illustrate with a diagram.

(6)

(ii) Write down a linear model to explain the yield y on a unit plot (i.e. a sub-plot). State and explain the properties of each of the terms in it.

(2)

(iii) List the items in the analysis of variance table for this experimental design, and state their degrees of freedom.

(2)

(iv) Explain briefly why the variance of the difference between the mean yields of two nitrogen levels is $2(\sigma_S^2 + 5\sigma_M^2)/15$, where σ_S^2 is the sub-plot component of variation and σ_M^2 is the main-plot component of variation. Explain also why the variance of the difference between the overall mean yields of two growth regulators is $\sigma_S^2/6$, and why the variance of the difference between two growth-regulator means at the *same* level of nitrogen is $2\sigma_S^2/3$.

(4)

(v) The main-plot residual mean square in an analysis of yield data (kg/plot) from this experiment is 24.47, the sub-plot residual mean square is 2.87 and the interaction between nitrogen and regulators is significant at the 5% level. Use the following table of means to examine the results of the experiment.

(6)

Growth Regulator	*G 1*	*G 2*	*G 3*	*G 4*	*G 5*	Overall *N* mean
Nitrogen level 1	11.5	14.8	17.4	11.8	14.3	(14.0)
Nitrogen level 2	14.0	15.8	20.7	18.2	15.3	(16.8)
Nitrogen level 3	12.1	15.3	19.0	14.9	17.2	(15.7)
Nitrogen level 4	13.3	12.0	18.2	15.1	16.7	(15.1)
Overall *G* mean	(12.7)	(14.5)	(18.8)	(15.0)	(15.9)	

E4. In a bioassay, random samples of individuals from a population are subjected to stimuli $\{d_i\}$ (which are doses of a compound), and the proportion p_i reacting to stimulus d_i in sample i is recorded.

(i) (a) Explain the term *tolerance distribution* as used in bioassay.

(b) The tolerance U in the population is a random variable having probability density function $f(u)$, $-\infty < u < \infty$. Write down an expression for the probability, π_i, of an individual reacting to stimulus d_i.

(c) Hence show that if U follows the logistic distribution, in which

$$f(u) = \frac{\exp\left[(u-\mu)/\tau\right]}{\tau\left\{1+\exp\left[(u-\mu)/\tau\right]\right\}^2}, \qquad \tau > 0, \quad -\infty < \mu < \infty,$$

then $\text{logit}(\pi_i) = \beta_0 + \beta_1 d_i$, where $\beta_0 = -\mu/\tau$ and $\beta_1 = 1/\tau$. (7)

(ii) (a) Under what conditions would it be better to measure stimulus as log dose, rather than dose, when analysing a set of data?

(b) Suppose that a series of stimuli $\{d_i\}$ has been used in an assay, and the observed values of π_i were p_i. The linear logistic model

$$\text{logit}(p) = \beta_0 + \beta_1 \log d$$

has been fitted, and estimates $\hat{\beta}_0$, $\hat{\beta}_1$ of β_0, β_1 have been found.

Derive estimates of $ED50$, the median effective stimulus (to which 50% of the population will react), and $ED90$ (to which 90% of the population will react).

(6)

(iii) (a) You may assume the (approximate) result that when $g(\hat{\beta}_0, \hat{\beta}_1)$ is a function of two parameter estimators, its variance is given by

$$\text{Var}(g) \approx \left(\frac{\partial g}{\partial \hat{\beta}_0}\right)^2 \text{Var}(\hat{\beta}_0) + 2\left(\frac{\partial g}{\partial \beta_0}\frac{\partial g}{\partial \hat{\beta}_1}\right)\text{Cov}(\beta_0, \hat{\beta}_1) + \left(\frac{\partial g}{\partial \beta_1}\right)^2 \text{Var}(\beta_1).$$

Show that, when log dose is used as explanatory variable,

$$\text{Var}(\log \hat{E}D50) \approx \left(\hat{v}_{00} - 2\frac{\hat{\beta}_0}{\hat{\beta}_1}v_{01} + \frac{\beta_0^2}{\beta_1^2}v_{11}\right)/\hat{\beta}_1^2,$$

where $\hat{v}_{00}$, v_{01} and $\hat{v}_{11}$ are functions which you should identify.

(b) Extracts from the computer output of an analysis of the effects of a certain serum (doses in cc) are as follows.

```
constant     -9.19          variance–covariance matrix
log dose     -1.83          constant     1.5753    0.3158
                            log dose     0.3158    0.0648
                                       constant  log dose
```

Construct an approximate 95% confidence interval for the $ED50$ of this serum.

(7)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. (i) At the end of each working day, a paint manufacturer measures the opacity of paint by selecting a sample of three tins at random from the day's production. The measurement is a computer camera assessment of how well a black and white check pattern on a cardboard test piece has been obliterated by a single coat. The scale runs from 0, transparent, to 100, obscured, and the target value is 94. Over the past 30 days, the average of the within-sample variances is 0.5662. The variance of the 30 means is 0.3486. The variance of all 90 data values is 0.7225.

(a) Estimate the within-sample standard deviation using the average within-sample variance. Explain why this is different from the mean of the 30 within-sample standard deviations. Which estimate would you prefer to use, and why?

(3)

(b) Estimate the between-sample variance and between-sample standard deviation. Hence estimate the standard deviation of opacity of paint in tins randomly selected from this process.

(3)

(c) Compare your estimate of the standard deviation of opacity of paint in tins randomly selected from this process with the standard deviation of all 90 data values. In general, would you have any reservations about estimating the overall variance of a process by calculating the variance of nk measurements made on k samples of size n?

(4)

(ii) You have been asked to set up two-sided Shewhart mean and range charts for the process described in part (i). Explain your construction of the charts, and demonstrate their use with the following data from 4 consecutive days of production. You may assume factors for the process standard deviation to give lower and upper action lines (0.001) on the range chart for samples of size 3 are 0.06 and 5.06 respectively.

Day 1	*Day 2*	*Day 3*	*Day 4*
92.18	91.60	93.76	91.93
93.32	91.96	91.93	91.28
92.33	92.07	92.84	92.28

(7)

(iii) The specification for opacity is that it should be between 91.5 and 96.5. Comment on the capability of the process.

(3)

F2. An experiment was performed on an integrated circuit device. Its objective was to reduce the variability of depth of an epitaxial layer whose target value was 14.5 microns. The control factors are: susceptor rotation method (A); nozzle position (B); deposition temperature (C); and deposition time (D). Six replicates of a full 2^4 factorial experiment were carried out. The results of the experiment are shown **at the bottom of this page and on the next page**. These contain a summary of a regression analysis with the means of the six replicates, at each factor combination, as the response, and a summary of a regression analysis with the natural logarithm of the variance of the six replicates, at each factor combination, as the response.

(i) (a) Using only the p-values printed in the regression summary, identify the factors you consider to have an effect on the mean depth. (You should use a nominal 10% significance level.) Calculate the coefficient of determination (R^2).

(b) Use the same method to identify the factors which have an effect on the variability.

(c) What values would you recommend setting for the control factors?

(d) What reservations do you have about the nominal p-values? How might you improve the analysis to allow for this?

(9)

(ii) An alternative experimental strategy would have been 12 replicates of a 2^{4-1} design.

(a) What would the advantages and disadvantages of this be?

(b) Now suppose that two noise variables E and F had also been identified, and that these noise variables could be held at fixed levels (–1) and (+1) in the experiment although they would vary randomly about 0 in routine production. Also the 12 replicates were 3 replicates of a 2^2 factorial design for E and F. One possible analysis is to regress the depths from each replicate (Y_i) on x_1, x_2, …, x_6, representing A, B, …, F respectively, and certain products of these. Write down a suitable model and explain how it might be used to achieve the objective.

(11)

Row	A	B	C	D	meandepth	vardepth
1	–1	–1	–1	1	14.821	0.003
2	–1	–1	–1	–1	13.860	0.005
3	–1	–1	1	1	14.757	0.003
4	–1	–1	1	–1	13.880	0.001
5	–1	1	–1	1	14.888	0.003
6	–1	1	–1	–1	14.165	0.004
7	–1	1	1	1	14.921	0.016
8	–1	1	1	–1	14.037	0.002
9	1	–1	–1	1	14.932	0.215
10	1	–1	–1	–1	13.972	0.121
11	1	–1	1	1	14.415	0.206
12	1	–1	1	–1	13.907	0.226
13	1	1	–1	1	14.878	0.147
14	1	1	–1	–1	14.032	0.088
15	1	1	1	1	14.843	0.327
16	1	1	1	–1	13.914	0.070

Results continued on next page

Regression analysis: meandepth versus *A*, *B*, *C*, *D*, *AB*, *AC*, *AD*, *BC*, *BD*, *CD*

The regression equation is

meandepth = 14.4 – 0.0273*A* + 0.0709*B* – 0.0546*C* + 0.4180*D* – 0.0158*AB* – 0.0373*AC* – 0.0126*AD* + 0.0236*BC* + 0.0048*BD* – 0.0183*CD*

Predictor	Coef	SE Coef	T	P
Constant	14.3889	0.03320	433.34	0.000
A	–0.02725	0.03320	–0.82	0.449
B	0.07087	0.03320	2.13	0.086
C	–0.05463	0.03320	–1.65	0.161
D	0.41800	0.03320	12.59	0.000
AB	–0.01575	0.03320	–0.47	0.655
AC	–0.03725	0.03320	–1.12	0.313
AD	–0.01262	0.03320	–0.38	0.719
BC	0.02362	0.03320	0.71	0.509
BD	0.00475	0.03320	0.14	0.892
CD	–0.01825	0.03320	–0.55	0.606

Analysis of variance

Source	DF	SS	MS	F	P
Regression	10	2.97892	0.29789	16.89	0.003
Residual Error	5	0.08820	0.01764		
Total	15	3.06712			

Regression analysis: 2log(s) versus *A*, *B*, *C*, *D*, *AB*, *AC*, *AD*, *BC*, *BD*, *CD*

The regression equation is

2log(s) = –7.54 + 3.830*A* + 0.092*B* + 0.066*C* + 0.615*D* – 0.444*AB* + 0.223*AC* + 0.020*AD* + 0.322*BC* + 0.347*BD* + 0.542*CD*

Predictor	Coef	SE Coef	T	P
Constant	–7.5432	0.2898	–26.03	0.000
A	3.8330	0.2898	13.23	0.000
B	0.0919	0.2898	0.32	0.764
C	0.0655	0.2898	0.23	0.830
D	0.6145	0.2898	2.12	0.087
AB	–0.4441	0.2898	–1.53	0.186
AC	0.2227	0.2898	0.77	0.477
AD	0.0197	0.2898	0.07	0.949
BC	0.3224	0.2898	1.11	0.317
BD	0.3470	0.2898	1.20	0.285
CD	0.5422	0.2898	1.87	0.120

Analysis of variance

Source	DF	SS	MS	F	P
Regression	10	253.561	25.356	18.87	0.002
Residual Error	5	6.719	1.344		
Total	15	260.280			

Turn over

F3. (a) Cathode ray tubes for a particular make of TV have an exponential lifetime distribution, regardless of whether or not they are being used. The mean lifetime of tubes is 7.5 years. Assume that all the tubes referred to below are selected randomly.

(i) What is the probability that a tube lasts at least 7.5 years?

(ii) What is the probability that a tube lasts a further 7.5 years if it has reached 7.5 years and is still working?

(iii) Suppose you have just bought one of these TVs and a spare tube. What is the probability that you have a TV with a working tube in 15 years' time given that the spare tube has the same lifetime distribution as the original tube? *Note that the spare tube may already have failed when you need to use it.*

(iv) Suppose you have just bought one of these TVs in a special promotion with a guarantee of one working replacement tube. What is the probability that you have a TV with a working tube in 15 years' time if the replacement tube has the same lifetime distribution as the original?

(v) You are now told that the manufacturer put a large number of tubes into a store at the time of the promotion. Assume your replacement tube would come from this store. Explain whether or not this changes your answer to part (iv).

(9)

Part (b) of this question is on the next page

(b) (i) Explain what is meant by the structure function ($\phi(\mathbf{x})$) for a system. A *k*-out-of-*n* system is one which will operate if and only if at least *k* of the *n* components operate. Its structure function can be expressed as

$$\phi(\mathbf{x}) = \min\left(1,\ \text{int}\left[\sum_{i=1}^{n} x_i / k\right]\right)$$

where int[.] is the integer value of a real number (e.g. int[7.8] = 7). Verify this form of the structure function for a 3-out-of-4 system.

(ii) A *k*-out-of-*n* system can be represented by a parallel system with duplicate nodes. For example, a 2-out-of-3 system can be represented by three parallel sub-systems, the sub-systems being: components 1 and 2 in series; components 1 and 3 in series; and components 2 and 3 in series. Give a graphical representation of a 3-out-of-4 system and hence find its structure function in an alternative form.

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

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

Use this definition to find the dual of two elements in parallel. What is the dual of two elements in series?

(iv) Give a graphical representation of the dual of the 3-out-of-4 system. Write down the structure function for this dual system, and give its alternative form.

(11)

F4. (i) A factory has three identical machines. The working times before failure for any one machine are independently exponentially distributed with a mean of $1/\lambda$ hours. There is one repair man. Repair times are independently exponentially distributed with a mean of $1/\theta$ hours.

Taking the state of this system as the number of working machines, what is the transition intensity matrix? Hence find the proportions of time that 0, 1, 2 and 3 machines are working. If the ratio $\theta/\lambda = 5$, what is the expected number of machine hours lost per hour?

(9)

(ii) A factory has three identical machines. The working times before failure for any one machine are independently exponentially distributed with a mean of $1/\lambda$ hours. Each machine has its own operator, and if it fails it is repaired by its operator. Repair times are independently exponentially distributed with a mean of $1/\rho$ hours.

Give the transition intensity matrix for this situation, and hence find the proportions of time that 0, 1, 2 and 3 machines are working. What is the expected number of machine hours lost per hour if (a) $\rho/\lambda = 5$, and (b) $\rho/\lambda = 3.5$?

(8)

(iii) What practical conclusions would you draw from your calculations of the expected number of machine hours lost per hour in cases (i) and (ii)?

(3)