EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA IN STATISTICS, 2008

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*

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

©RSS 2008

SECTION A – STATISTICS FOR ECONOMICS

A1. Three administrative districts A, B and C in a large town consist of $N_A = 5000$, $N_B = 3000$ and $N_C = 2000$ households respectively. A sociologist wishes to carry out a 10% sample survey of households to estimate the total number of resident children in these districts.

(i) A sample of 1000 households is taken (without replacement) from the total number of 10000 households, and the number of resident children x_i is recorded in each sampled household. The sum of these observations is $\Sigma x_i = 1429$ and the sum of their squares is $\Sigma x_i^2 = 4236$. Assuming that the data are a simple random sample from all 10000 households, obtain an approximate 95% confidence interval for the total number of children in districts A, B and C. State any assumptions needed for your analysis.

(6)

(ii) Suppose now that the sample is in fact a stratified random sample with districts as strata, allocated proportionally to the number of households in each district, and that the sum and sum of squares of the observations in each district are as shown in the following table.

District	*Number of households sampled*	$\sum x_i$	$\sum x_i^2$
A	$n_A = 500$	556	1253
B	$n_B = 300$	426	1179
C	$n_C = 200$	447	1804

Obtain an approximate 95% confidence interval for the total number of children in districts A, B and C, based on the proportionally allocated stratified sample, making your reasoning clear.

(9)

(iii) Compare the results of your analyses in parts (i) and (ii). Why is stratified sampling usually expected to yield more precise results than a simple random sample of the same size? If your estimated within-stratum sample variances in part (ii) were in proportion to the true within-stratum variances (and unit costs of sampling were assumed to be equal in all strata), what would be the optimal allocation of sampling effort across the three districts? Comment on your results in the light of this allocation.

(5)

A2. (i) Briefly explain the purpose of *standardisation* in comparing mortality rates across (for example) diverse regions or occupational groups in a country.

(5)

(ii) Coal miners in a developing country are classified into three groups: high risk (H), medium risk (M) and low risk (L). The table below shows numbers at risk and 10-year numbers of deaths in each group, nationally and in two regions of the country, A and B.

		Risk group		
		H	M	L
National	*At risk*	10000	30000	60000
	Deaths	1300	1500	1200
A	*At risk*	800	1200	3000
	Deaths	96	60	54
B	*At risk*	200	1000	2800
	Deaths	28	60	56

(a) Calculate crude and standardised 10-year death rates for each of regions A and B using the direct method of standardisation, making your reasoning clear. Comment briefly on your results.

(5)

(b) Suppose now that at least some of the regional numbers of deaths within each group are considered to be too small to provide sufficiently reliable death rates. Explain what is meant by the *index death rate* for a region, and calculate this quantity for each region. For each region, compare the index death rate with the national rate and comment.

(5)

(c) Use the crude regional death rates in conjunction with the index death rates obtained in part (b) to obtain regional standardised mortality ratios, and comment briefly on your results.

(5)

A3. Investment trusts are companies whose sole activity is owning shares of other companies. The dividend paid by an investment trust is the total of the dividends which it receives from the companies whose shares it holds, less the costs of running the trust.

Random samples of 13 investment trusts and 20 other companies are taken and their dividend yields (%), arranged in ascending order, are as follows.

Investment trusts	0.0 0.1 0.3 0.7 1.1 1.5 1.6 1.9 2.6 3.4 4.5 5.6 7.9	$\Sigma x = 31.2$ $\Sigma x^2 = 142.56$
Other companies	0.9 1.2 1.7 2.2 2.5 3.3 3.8 3.9 4.1 4.3 4.8 5.2 6.0 6.4 6.5 7.1 8.0 8.8 9.1 10.2	$\Sigma x = 100.0$ $\Sigma x^2 = 640.06$

(i) (a) It is required to test the null hypothesis that the variances in the two populations are equal. Justify an economically appropriate alternative hypothesis, carry out the test, and comment on the result from an economist's standpoint.

(6)

(b) Assuming that the variances may be taken as equal, test the null hypothesis that the means in the two populations are equal against a two-sided alternative. Suggest two possible economic reasons for the observed difference between the two population means.

(7)

(ii) Comment critically (but without further calculations) on the validity of the assumptions necessary for carrying out the test in part (i)(b). Briefly describe, but do not perform, a non-parametric test which could be applied to these data. State carefully the relevant null and alternative hypotheses for this test. Would you regard the non-parametric test applied to these data as having equal, more, or less validity compared with the test in part (i)(b)? Why?

(7)

A4. It is required to estimate an Engel expenditure curve relating household expenditure on recreation and culture, y, to total household expenditure, x. A commonly used model for fitting a sample of expenditures (x_i, y_i), $i = 1, 2, \ldots, n$, is

$$y_i = \beta_1 + \frac{\beta_2}{x_i} + e_i \qquad (\beta_1 > 0,\ \beta_2 < 0),$$

where the e_i are assumed to be independent, identically Normally distributed error terms with mean zero.

(i) Briefly explain the reasoning behind this model, and its limitations. What is the economic significance of the quantities β_1 and $-\beta_2/\beta_1$?

(4)

The data below are compiled from published official statistics for the period 2001–2002.

Average Weekly Household Expenditure by Gross Income Decile Group, 2001–2002 (£)

	y (£)	x (£)	$1/x$
1st decile (lowest 10%)	15.5	97.5	0.0102564
2nd decile	21.8	99.9	0.0100100
3rd decile	26.0	114.2	0.0087566
4th decile	39.0	133.1	0.0075131
5th decile	43.4	139.5	0.0071685
6th decile	49.4	148.9	0.0067159
7th decile	63.9	164.9	0.0060643
8th decile	76.6	184.0	0.0054348
9th decile	84.5	199.0	0.0050251
10th decile (highest 10%)	119.8	281.8	0.0035486

(ii) Plot a graph of the data with a view to assessing the suitability of an Engel model of the above form, and comment briefly.

(5)

(iii) Obtain ordinary least squares (OLS) estimates of the parameters β_1 and β_2 for the Engel model of the above form. For this purpose you are given that

$\Sigma y_i = 539.9$, $\Sigma(1/x_i) = 0.0704933$, $\Sigma(1/x_i^2) = 0.000539166$, $\Sigma(y_i/x_i) = 3.19431$.

Provide a point estimate of the expected expenditure on recreation and culture given a total weekly expenditure of £150. Also give a point estimate of the threshold level of total expenditure below which nothing is spent on recreation and culture.

(5)

Question A4 is continued on the next page

(iv) Summary information from OLS fitting of the extended Engel model

$$y_i = \beta_1 + \frac{\beta_2}{x_i} + \frac{\beta_3}{x_i^2} + e_i$$

is as follows.

Predictor	Coef	SE Coef	T	P
Constant	243.181	7.060	34.45	0.000
1/x	−41092	2070	−19.85	0.000
1/x_sq	1863646	143441	12.99	0.000

S = 1.95696 R-Sq = 99.7% R-Sq(adj) = 99.6%

Signs of standardized residuals: − + − − − − + + + − .

Briefly explain with reasons whether or not you consider the extended model to be satisfactory as a summary of these data. Use the extended model to estimate the expected expenditure on recreation and culture given a total weekly expenditure of £150, and compare this answer with the corresponding result of part (iii).

Comment critically on the possibility of using the extended model outside the range of the available data.

(6)

Turn over

SECTION B – ECONOMETRICS

B1. The time series $\{y_t\}$ consists of the annual income in £ to the Receiver of Seamen's Sixpences for the years 1717–1828 inclusive. For $t = 1, 2, \ldots, 112$, the series $\{x_t\}$ is defined by $x_t = \log y_t$, and for $t = 2, 3, \ldots, 112$, the series $\{z_t\}$ is defined by $z_t = x_t - x_{t-1}$.

The edited computer output **on the next page** provides some summary analyses of the series $\{z_t\}$ and $\{x_t\}$.

(a) (i) Identify the model and write down its equation, for each of the statistical Models 1, 2 and 3 fitted to the series $\{z_t\}$.

(3)

(ii) Discuss with reasons which of the above fitted models for the series $\{z_t\}$ you would choose.

(7)

(iii) Briefly specify any further diagnostic checks you might wish to make on your chosen model.

(2)

(b) (i) Write down the equation for Model 1 in terms of the untransformed series $\{y_t\}$. Given that log(income) for 1828 is $x_{112} = 9.92220$ and that the estimated residual for 1828 ($t = 112$) is $\hat{\varepsilon}_{112} = 0.005306$, calculate point forecasts for the y values in the years 1829 and 1830, based on this model.

(5)

(ii) Estimate the annual percentage increase in the y values based on the fitted Model 1.

(3)

The edited computer output is on the next page

Autocorrelation Function of $\{z_t\}$

Lag	1	2	3	4	5	6	7	8
Corr	−0.412	−0.053	0.033	0.062	−0.097	0.064	−0.052	−0.010

It may be assumed that no further correlations exceed 0.1 in absolute value.

Model 1 ARIMA 0 1 1 'X'

Type	Coef	SE Coef	T	P
MA 1	0.5893	0.0774	7.62	0.000
Constant	0.010136	0.004323	2.34	0.021
(Intercept)				

Differencing: 1 regular difference
Number of observations: Original series 112, after differencing 111
Residual Mean Square = 0.01219 DF = 109

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6.5	12.5	18.5	28.9
DF	10	22	34	46
P-Value	0.770	0.947	0.986	0.977

Model 2 ARIMA 1 1 0 'X'

Type	Coef	SE Coef	T	P
AR 1	−0.4126	0.0873	−4.73	0.000
Constant	0.01387	0.01099	1.26	0.209
(Intercept)				

Differencing: 1 regular difference
Number of observations: Original series 112, after differencing 111
Residual Mean Square = 0.01340 DF = 109

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	13.5	21.6	28.3	43.2
DF	10	22	34	46
P-Value	0.198	0.485	0.742	0.591

Model 3 ARIMA 1 1 1 'X'

Type	Coef	SE Coef	T	P
AR 1	0.0873	0.1586	0.55	0.583
MA 1	0.6565	0.1201	5.47	0.000
Constant	0.009329	0.003630	2.57	0.012
(Intercept)				

Differencing: 1 regular difference
Number of observations: Original series 112, after differencing 111
Residual Mean Square = 0.01228 DF = 108

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6.6	12.0	17.9	28.7
DF	9	21	33	45
P-Value	0.683	0.939	0.985	0.972

Turn over

B2. An OLS (ordinary least squares) regression model is sought for the sales of Quenchers Soft Drinks on the basis of quarterly observations over the 5-year period 2003–2007, using time t and other predictor variables as explained below.

In the edited summary computer output tabulated below, Y denotes $\log_{10}$(sales), where sales are in units of £10000; data are quarterly from Winter (January February March), Spring (April May June), Summer (July August September), Autumn (October November December); and time t runs from Winter 2003 = 1 to Autumn 2007 = 20. For i = 1, 2, 3, 4, four quarterly dummy variables Q_i are defined to equal 1 in the ith quarter and zero otherwise. The explanatory variable *seas1* is defined as $-Q_1 + Q_2 + Q_3 - Q_4$, and the explanatory variable *seas2* is defined as $Q_3 - Q_4$. Estimated standard errors of the regression coefficients are given in parentheses below the coefficients to which they refer. Four fitted models, (1), (2), (3), (4), are shown in the table.

$\hat{Y}$ = 2.36730 + 0.018784t (0.04967) (0.004147) R^2 = 0.517, s =0.10693, Resid ss = 0.20581 DW = 2.457	(1)
$\hat{Y}$ = 2.33446 + 0.018791t + 0.05092Q_2 + 0.13766Q_3 − 0.08270Q_4 (0.04125) (0.002853) (0.04574) (0.04601) (0.04645) R^2 = 0.816, s = 0.0721847, Resid ss = 0.078159 DW = 2.104	(2)
$\hat{Y}$ = 2.36730 + 0.018184t + 0.06782*seas1* (0.03801) (0.003173) (0.01830) R^2 = 0.733, s = 0.0818258, Resid ss = 0.11382 DW = 2.827	(3)
$\hat{Y}$ = 2.35859 + 0.019013t + 0.11029*seas2* (0.03282) (0.002741) (0.02235) R^2 = 0.801, s = 0.0705562, Resid ss = 0.08463 DW = 2.084	(4)

(i) Test each of Q_2, Q_3 and Q_4 in Model 2 for partial significance and explain why only three of the four dummy variables are used in this model. What would you expect to find if Q_1 were also included in the regression?

(4)

(ii) Explain how you might use the above results to test for the seasonality of sales, and carry out this test.

(4)

(iii) Suggest reasons why models (3) and (4) were fitted.

(3)

(iv) Comment on the DW (Durbin-Watson) statistics for each of these regressions. Making any further tests you consider necessary, discuss which of the four models best represents the data.

(5)

(v) Using your chosen model of part (iv), give point forecasts of **actual** sales for each quarter of 2008.

(4)

Turn over

B3. The level of beer consumption in year t (y_t, in millions of bulk barrels) is modelled as a function of its relative price (x_{1t}) and the level of personal disposable income (x_{2t}, measured in £000 at 1988 prices). The model used is

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t ,$$

where the $\{ \varepsilon_t \}$ are taken to be independent Normally distributed random variables with mean 0 and variance σ^2.

Data are available for the 17 years from 1988 to 2004 inclusive and give the following summary quantities.

$$\sum_{t=1}^{17} y_t = 612 \qquad \sum_{t=1}^{17} (y_t - \bar{y})^2 = 415 \qquad \sum_{t=1}^{17} x_{1t} = 1411 \qquad \sum_{t=1}^{17} (x_{1t} - \bar{x}_1)^2 = 2100$$

$$\sum_{t=1}^{17} x_{2t} = 1139 \qquad \sum_{t=1}^{17} (x_{2t} - \bar{x}_2)^2 = 1272 \qquad \sum_{t=1}^{17} (x_{1t} - \bar{x}_1)(y_t - \bar{y}) = -859$$

$$\sum_{t=1}^{17} (x_{2t} - \bar{x}_2)(y_t - \bar{y}) = 682 \qquad \sum_{t=1}^{17} (x_{1t} - \bar{x}_1)(x_{2t} - \bar{x}_2) = -1521$$

(i) Find the least squares estimates of β_1 and β_2 and their estimated standard errors.
(6)

(ii) Test the hypothesis H_0: $\beta_1 = 0$ against an economically plausible alternative.
(2)

(iii) Test the hypothesis H_0: $\beta_2 = 0$ against an economically plausible alternative.
(2)

(iv) Test the hypothesis H_0: $\beta_1 = \beta_2 = 0$ against the alternative hypothesis that at least one of β_1 and β_2 is non-zero.
(2)

(v) Test the hypothesis H_0: $2\beta_1 + \beta_2 = 0$ against H_1: $2\beta_1 + \beta_2 \neq 0$.
(5)

(vi) Comment briefly on the results of your tests.
(3)

B4. (a) (i) What is meant by *heteroscedasticity* in an econometric regression model?

(2)

(ii) What are the consequences for OLS (ordinary least squares) estimators in a regression model if heteroscedasticity is present?

(2)

(iii) Outline **two** approaches for dealing with heteroscedasticity in a regression model.

(4)

(b) Summary results from OLS regression analysis of the relationship between consumption expenditure (y) and net income (x), both in \$ per day, are tabulated below.

Regression	Model	R^2	Resid SS
1	$y = a + bx + error$ (full regression, $n = 30$)	0.947	2361.0
2	$y = a + bx + error$ ($n = 10$, lowest 10 incomes)	0.785	292.9
3	$y = a + bx + error$ ($n = 10$, highest 10 incomes)	0.708	1204.0
4	$\hat{e}^2 = a + bx + error$ $\quad (n = 30)$ ($\hat{e}$ = estimated error from regression 1)	0.176	302972.0

Each of the tests Goldfeld-Quandt (GQ) and Breusch-Pagan (BP) may be used to test the null hypothesis of homoscedasticity against the contrary alternative hypothesis. GQ uses the usual F test for equality of variances, applied to the residual variances from fitting the same regression to two separate subsets of the data, with a significant result indicating rejection of the null hypothesis. BP regresses the squares of the residuals from the fitted model on the regressor variables (see regression 4 above), and takes as test statistic the regression sum of squares divided by twice the (ML-estimated) residual mean square from the original regression. Under the null hypothesis, this statistic has an approximate χ^2 distribution with degrees of freedom equal to the number of explanatory variables used.

Assess whether heteroscedasticity is present in the data as modelled by Regression 1 by applying **each** of

(i) the Goldfeld-Quandt test, (3)

(ii) the Breusch-Pagan test. (5)

Comment briefly on your findings, including reliance on any assumptions unsupported by the summary results above.

(4)

SECTION C – OPERATIONAL RESEARCH

C1. (a) Consider a machine that is to operate continuously for the foreseeable future. Its maintenance cost is at per unit time, where t is its age. At regular intervals of length T the machine is replaced with a new machine of the same type.

Suppose that the cost of replacing the machine is K. This includes any costs due to downtime. Derive the average cost per unit time for operating the machine and thus show that the optimal value of T is $\sqrt{2K/a}$.

(8)

(b) The lifespan of a certain machine has probability density function

$$f(t) = t^{-2}, \quad t > 1.$$

The machine can be replaced before it breaks down at a cost of £10 000, or after breakdown at a cost of £15 000.

Suppose that we replace the machine at age T if it has not yet broken down. Calculate the expected cost per lifecycle and expected length of lifecycle for the machine. (Note that we are considering only replacement costs, not maintenance costs.)

(6)

At what age should the machine be replaced?

(6)

C2. (a) Consider a Markovian queue with a buffer of size 3 (that is, the maximum number waiting to begin service is 3) and two servers, A and B. Customers arrive at rate λ, server A serves at rate α and server B at rate β. If only one customer is present then server A is used (a customer switching immediately from server B if necessary); if more customers are present then both servers are used.

(i) Draw a transition diagram for this system, clearly indicating the transition rates. Use the numbers of people in the system as the states.

(3)

(ii) Let $\pi(i)$ be the limiting probability of being in state i. Write down and solve the detailed balance equations for the $\pi(i)$.

(5)

(b) A dock company owns one berth with a crane, and further space for up to two ships to wait for unloading. Ships arrive at a rate of four per week, and any ships that arrive while all three spaces are full are lost to rival docks. It takes on average two days to unload a ship with the crane, and the dock company earns £10 000 per day for its use. The dock works 52 weeks each year.

(i) The company is considering converting one of the two spare berths into an unloading bay, by installing a second similar crane at a cost of £5 000 000. Using Markovian queuing models, in which the states are the numbers of ships in the dock, draw transition diagrams for both the existing and proposed models. By solving these systems, estimate how many years it would take to pay off this expense from the additional profit.

(10)

(ii) Which of the assumptions implicit in your model do you consider the least appropriate, and why?

(2)

C3. (i) Formulate the following goal programming problem as a Lexicographic/Pre-emptive linear programming problem.

$$\begin{array}{lrcl} \text{Goal 1} & x_1 + x_2 & \geq & 6 \\ \text{Goal 2} & x_1 + 2x_2 & \geq & 10 \\ \text{Subject to} & x_1 - x_2 & \leq & 1 \\ & x_1, x_2 & \geq & 0 \end{array}$$

(4)

(ii) Solve the problem, assuming Goal 1 is the more important.

(16)

Turn over

C4. For both parts of this question, assume that an ample supply of independent U(0, 1) random variables U_1, U_2, U_3, … is available.

(a) Suppose that X_1, X_2, … are independent random variables, all having the logistic density function

$$f(x) = \frac{\exp\left(-\dfrac{x-\alpha}{\beta}\right)}{\beta\left(1+\exp\left(-\dfrac{x-\alpha}{\beta}\right)\right)^2} \quad \text{for } x \in (-\infty, \infty),$$

where α and β are known parameters. Also suppose that N is independent of the X_i and has a Poisson distribution with mass function

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x = 0, 1, \ldots .$$

Let $Y = \sum_{i=1}^{N} X_i$. Explain how you would use U_1, U_2, U_3, … to simulate from the distribution of Y. Give pseudo-code describing your algorithm.

(12)

(b) Suppose that you have a closed form expression for the density function $g(x)$ of X that satisfies

$$g(x) \begin{cases} \leq \alpha\lambda e^{-\lambda x} & \text{for } x \geq 0 \\ = \quad 0 & \text{otherwise} \end{cases}$$

where α and λ are known positive constants, but no closed form expression for either the cumulative distribution function G or its inverse G^{-1}.

Explain how you would use U_1, U_2, U_3, … to simulate from the distribution of X. Give pseudo-code describing your algorithm.

(8)

SECTION D – MEDICAL STATISTICS

D1. Researchers wished to compare two different methods, SBB and Control, of delivering analgesia during shoulder surgery. They were interested in whether post-operative pain experience was better with the new therapy (SBB) compared with the old therapy (Control), so they treated 19 randomly chosen patients with SBB and another 15 patients with Control.

(i) (a) What is meant by a two-sided hypothesis? Write down suitable null and alternative hypotheses for this study.

(3)

(b) Two-sided alternative hypotheses are often used except when the study is a non-inferiority study or an equivalence study.

What is meant by *non-inferiority* and by *equivalence*?

How are the sample size calculations for standard comparative trials affected when these two types of study are carried out?

(5)

Post-operative pain was measured using visual analogue scores recorded at 1, 2, 4, 8, 12 and 24 hours post-operative. These measurements were combined by calculating the area under the curve (AUC) and this was used as a suitable summary measure of the average effect for post-operative pain. The table shows the mean and standard deviation of AUC for each group, and also the median and sample size.

		AUC			
		N	*Mean*	*Median*	*St Dev*
Group	*SBB*	19	39.7	36.5	23.9
	Control	15	67.0	67.5	20.2

(ii) Stating any assumptions that you make, perform an appropriate hypothesis test to compare the mean AUC between the SBB and Control therapy groups. Calculate a 95% confidence interval for the difference in the mean AUC between the two therapies. Comment on the inferences that can be deduced from this hypothesis test and confidence interval.

(8)

(iii) The researchers were concerned that age and sex might be important factors in determining post-operative pain. The output **on the next page** is from a multiple regression analysis of these data including group, age and sex in the model. Examining the output from this model, would you conclude that either age or sex were significant risk factors? Comment on your conclusions.

(4)

Results of regression analysis for question D1 part (iii)

Source	SS	df	MS		
Model	7946.85	3	2648.95	Number of obs	34
Residual	14323.76	30	477.46	F(3, 30)	5.55
Total	22270.62	33	674.87	Prob > F	0.0038
				R-squared	0.3568
				Adj R-squared	0.2925
				Root MSE	21.851

auc	Coef	Std Err	t	P>|t|	95% Conf	Interval
group	26.95	7.56	3.57	0.001	11.52	42.38
age	0.54	0.30	1.82	0.079	-0.07	1.14
sex	1.53	7.68	0.20	0.844	-14.16	17.22
intercept	-44.18	24.44	-1.81	0.081	-94.09	5.72

Turn over

D2. (i) Define the following terms in respect of an event occurring.

(a) Absolute risk

(b) Relative risk

(c) Odds

(d) Odds ratio.

(4)

(ii) In what circumstances can an odds ratio be used to approximate to a relative risk? Explain why this is the case.

(2)

In a recent study conducted to examine why some cases of paediatric meningococcal disease were fatal whilst others were not, the researchers found that there was an absence of paediatric care in 33 of the 133 fatal cases, whilst there was an absence of paediatric care in 32 of the 355 non-fatal cases.

(iii) What type of study is this?

(1)

(iv) Construct an appropriate 2×2 contingency table showing the results given above. Calculate both a point estimate and a 95% confidence interval for the odds ratio of absence of care for fatal cases compared with non-fatal cases. Comment on the results.

(11)

(v) The researchers were also interested in whether other factors might be important in distinguishing between fatal and non-fatal cases. Some possible factors were severity of illness, sex and failure of management. How might one model the influence of these different variables?

(2)

D3. The table shows the time to revision surgery (survival time in years) for 17 patients who had had a previous hip replacement. The study investigators were interested in whether there was a difference in survival times for patients whose reason for revision surgery was aseptic loosening of the hip compared with other reasons.

Survival time for aseptic loosening group	*Survival time for others*
10.16	11.21
10.56	6.85
12.81	11.08
10.47	9.56 *
17.35 *	1.21
4.03	5.11
15.00	15.02 *
13.67	1.00
12.27	

* indicates right-censored observations

Explain what is meant by a right-censored observation.

(1)

Compute Kaplan-Meier survival curves for the two treatment groups and show both curves on one graph. Estimate the median time to revision for each group from this graph.

(10)

The log rank statistic for these data was 0.62 on 1 df, $p = 0.43$. Comment on this result. Explain whether you would expect the confidence interval for the difference in survival times to include or exclude 0.

(2)

Previous studies have suggested that in addition to aseptic loosening, age and sex could be important factors in determining the time to revision surgery. What method could you use to investigate whether these factors were indeed important? What is the key assumption for this method and how can it be tested?

(3)

On the next page is shown some abbreviated computer output from fitting two models to the data above, one using only aseptic loosening as a predictor, and one including age and sex in addition. Sex was coded as "male" = 1, "female" = 2. Aseptic loosening was coded as "no" = 0, "yes" = 1. Comment on the regression coefficients from this model.

(4)

Computer output for question D3

```
aseptic

         failure _d:  revision
   analysis time _t:  time_to0

No. of subjects =           17                     Number of obs   =       17
No. of failures =           14
Time at risk    =       167.36
                                                    LR chi2(1)      =     0.60
Log likelihood  =       -30.03                      Prob > chi2     =     0.44

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     aseptic |        .65          .36    -0.78   0.43        .22    1.91
------------------------------------------------------------------------------
```

```
aseptic age sex

         failure _d:  revision
   analysis time _t:  time_to0

No. of subjects =           16                     Number of obs   =       16
No. of failures =           13
Time at risk    =       166.15
                                                    LR chi2(3)      =     2.93
Log likelihood  =       -26.03                      Prob > chi2     =     0.40

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     aseptic |        .64          .38    -0.75   0.46        .20    2.05
         age |        .92          .05    -1.58   0.11        .83    1.02
         sex |        .71          .48    -0.51   0.61        .19    2.67
------------------------------------------------------------------------------
```


Turn over

D4. Describe the difference between the direct and indirect methods of standardisation of mortality rates. Comment on when each might be used.

(5)

The number of babies born varies across populations depending upon such factors as the size and age structure of the population. In order to make comparisons across populations it is necessary to standardise for any difference in the age distributions that might occur between populations. The table shows the number of women between the ages of 15 and 49 for the English cities Birmingham and Sheffield, together with the age-specific birth rates for England and Wales for 2005. In 2005 the actual number of live births was 15 893 for Birmingham and 6 101 for Sheffield. Assuming that all the births occurred to women between the ages of 15 and 49, what was the crude birth rate for women in this age range, in Birmingham in 2005? What was the rate in Sheffield? Compare your results for the two cities.

(3)

Age	*Age-specific fertility rates, England and Wales per 1 000 women, 2005*	*Birmingham population*	*Sheffield population*
15–19	26.3	35 789	16 491
20–24	71.2	38 530	20 351
25–29	98.8	35 264	17 143
30–34	100.9	38 316	19 538
35–39	50.3	36 815	19 205
40–44	10.3	31 452	17 145
45–49	0.6	27 892	14 911
	Total	244 058	124 784
	Total live births	15 893	6 101

Using the birth rate for England and Wales as a standard, calculate for each city (Birmingham and Sheffield) the expected number of births in 2005. Calculate also, for each city, the standardised birth ratio and a 95% confidence interval for this ratio.

(8)

Compare your results for the two cities.

(4)

SECTION E – BIOMETRY

E1. A research worker has carried out an experiment on a crop which is planted in rows in the field. The experimental units were plots of a fixed size, and four treatments A, B, C and D were compared. Two pieces of land, in adjacent fields, had to be used because each was only large enough to take a 4×4 Latin square layout. The experimenter has analysed the data on crop yield from each field separately, and is surprised that the results seem slightly different. He wants to write a report covering the whole experiment, and calls in the statistician to explain how to include a possible difference between fields (squares) in the full analysis.

The data, coded into convenient units, and useful sums of squares from his analysis, follow.

Field 1					
					Total
	A: 14	D: 7	B: 13	C: 18	52
	D: 8	A: 12	C: 17	B: 15	52
	B: 14	C: 17	D: 9	A: 10	50
	C: 16	B: 12	A: 8	D: 10	46
Total	52	48	47	53	200
Totals A: 44 B: 54 C: 68 D: 34					
Sums of squares: $14^2 + 7^2 + 13^2 + \ldots + 10^2 = 2690$ $52^2 + 52^2 + 50^2 + 46^2 = 10024$ $52^2 + 48^2 + 47^2 + 53^2 = 10026$ $44^2 + 54^2 + 68^2 + 34^2 = 10632$					

Field 2					
					Total
	D: 15	A: 17	C: 20	B: 20	72
	A: 20	C: 19	B: 21	D: 19	79
	B: 18	D: 13	A: 15	C: 18	64
	C: 23	B: 17	D: 12	A: 13	65
Total	76	66	68	70	280
Totals A: 65 B: 76 C: 80 D: 59					
Sums of squares: $15^2 + 17^2 + 20^2 + \ldots + 13^2 = 5050$ $72^2 + 79^2 + 64^2 + 65^2 = 19746$ $76^2 + 66^2 + 68^2 + 70^2 = 19656$ $65^2 + 76^2 + 80^2 + 59^2 = 19882$					

(i) Construct the analysis of variance for each field (square) separately.

(8)

(ii) Construct also a single analysis of variance for the whole experiment which includes a term for the difference between fields, deals in a suitable way with the terms for rows and columns within squares, and allows the treatments to be compared on the basis of their means over the whole experiment.

Comment on the results, and on the statistical assumptions made during the analysis.

(12)

E2. A field experiment using plots (units) each planted with 40 strawberry plants was laid out in 4 randomised blocks. There were 8 treatments A – H which are described in part (ii). The layout was as shown below, with North at the top of the diagram.

	A1	F1	D2	E2	G3	D3	G4	A4	↑ North
	G1	D1	B2	F2	B3	A3	D4	C4	
	H1	E1	H2	A2	C3	H3	B4	F4	
	B1	C1	C2	G2	E3	F3	E4	H4	
Block	I		II		III		IV		

Some plants failed to grow after planting, and the number of failures w on each plot was recorded. It was also noticed during the growing season that there were several poor plants on the North side of the plantation, but the actual number was not recorded.

(i) The total crop for the season on each plot, y, was recorded and the analysis of these data was to be made incorporating corrections for w and for the poor growth on the North side. Explain how a suitable indicator variable x could be constructed to correct for the poor growth, and state what values it would take on which plots.

Write down the linear model that would be the basis for the analysis.

(5)

(ii) Before planting, the plots either had cover crops growing on them to keep down weeds and improve soil quality (treatments A – G) or were ploughed and had no cover crop (treatment H). The cover crops were two different grasses A and B, two different fescues C and D, two different clovers E and F, and G was Lucerne Grass.

Suggest seven orthogonal contrasts among these treatments which could be tested in the analysis and which would give useful information to include in a report. State the coefficients required in each one.

Comment on whether all the useful contrasts that could be made would in fact be orthogonal.

(12)

(iii) Suppose that plot H4 was damaged by machinery during the season. How might this be allowed for in the analysis?

(3)

E3. The Agricultural Service in a large country is planning to carry out a survey of the production of a major ground crop. This has not been done for a few years but there are maps of all the main areas where the crop is grown and there is a budget for one aerial survey to be carried out early in the growing season. Some parts of the country are reasonably flat, while others are hilly; the crop can be grown in both conditions. The survey (or surveys) carried out on the ground should cover both conditions. Information is required on the area planted with this crop and on the yield of the crop when harvested.

Draw up a detailed proposal for this survey. Your proposal should include advice on the following topics, and on any others which you consider important.

- The use which should be made of an aerial survey, and how it may be combined with information obtained from surveys of the crop on the ground during the growing season
- The method of sampling, especially whether stratification and/or cluster sampling might be used
- Whether ratio or regression methods can be used to improve the estimates obtained from sampling the crop on the ground
- What supplementary data, for example use of fertiliser and method of cultivation (hand, mechanical, etc), it would be useful to collect
- How data should be collected so that transfer for analysis by computer may be made accurately

(20)

E4. Logistic functions are used in two major ways in biometry:

(a) in modelling the growth of plants or organisms over time;

(b) in the analysis of binary response data in assays.

For (a), explain the three phases of growth as size y moves from an initial small value towards a final approach to a limiting size as time x increases. Discuss how the logistic model arises, sketch the growth curve and describe how the parameters in the model may be estimated. Comment also on the problems that may arise in designing experiments to study growth, and how to choose times $\{x_i\}$ at which to measure size $\{y_i\}$.

(10)

For (b), explain the application of a (linear) logistic model in studying how proportions p_i depend on the values x_i of a single explanatory variable x. Discuss how the results of such a study should be presented and used in biological assay.

(10)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A company manufactures springs for fuel injectors. The customer's specification for length of each spring is that it should be between 49 mm and 51 mm. Assume that lengths of springs are independent and have a Normal distribution with mean 50 mm and standard deviation 0.30 mm.

(i) Calculate the process capability index. Use the *Statistical tables for use in examinations* to calculate a least upper bound for the parts per million outside the specification. (2)

(ii) Random samples of 5 springs are taken every shift, and the mean and range of the lengths of the sampled springs are calculated. The results from the first four shifts are as follows.

Shift	*Sample size*	*Mean (mm)*	*Range (mm)*
1	5	50.2	0.80
2	5	49.9	0.26
3	5	50.1	1.07
4	5	49.5	0.85

(a) Construct Shewhart mean and range charts. The mean chart should include both action and both warning lines. Report your conclusions. (6)

[You are given that the multiplying factors for the process standard deviation that give lower and upper 0.1% points for the range of a random sample of 5 from a Normal distribution are 0.37 and 5.48 respectively.]

(b) Draw a cumulative sum (cusum) chart for the means. (3)

(c) Use the following algorithm to decide whether action should be taken on the basis of your cusum chart. Let the target value be τ, the standard deviation of the variable that is being plotted be θ, and K be 0.5θ. Set $SH(0) = 0$ and $SL(0) = 0$. Calculate

$$SH(t) = \max[0,\ (x_t - \tau) - K + SH(t-1)]$$
$$SL(t) = \max[0,\ -(x_t - \tau) - K + SL(t-1)]$$

for $t = 1, 2, 3, 4$. Action is indicated if any of the $SH(t)$ or $SL(t)$ values exceeds 5θ. Comment on your calculations. (5)

(iii) Suppose the process mean shifts to 49.70 mm, with the standard deviation unchanged.

(a) Calculate the new process performance index. (2)

(b) What is the probability the next sample mean will be below the lower action line? What is the expected run length before action is indicated? (2)

F2. A company manufactures aluminium aircraft engine blocks using a casting process, and an engineer is asked to conduct an experiment that investigates the porosity of the blocks with the objective of determining conditions that minimize porosity. The engineer considers the 6 control factors described in the following table, each at two levels, in a single performance of a one-quarter fraction of the full factorial design. The response is porosity measured on a scale from 0 to 1000.

Factor	*Description*	*Coding for regression*
A	Mould lining compound	Type I (–1) Type II (+1)
B	Time in mould	Low (–1) High (+1)
C	Time of pour	Low (–1) High (+1)
D	Melt temperature	Low (–1) High (+1)
E	Quench temperature	Low (–1) High (+1)
F	Pressure applied to mould	Low (–1) High (+1)

(i) The design generators are ABC = E and BCD = F. Write down the aliases of A, of AB and of AE**.**

(5)

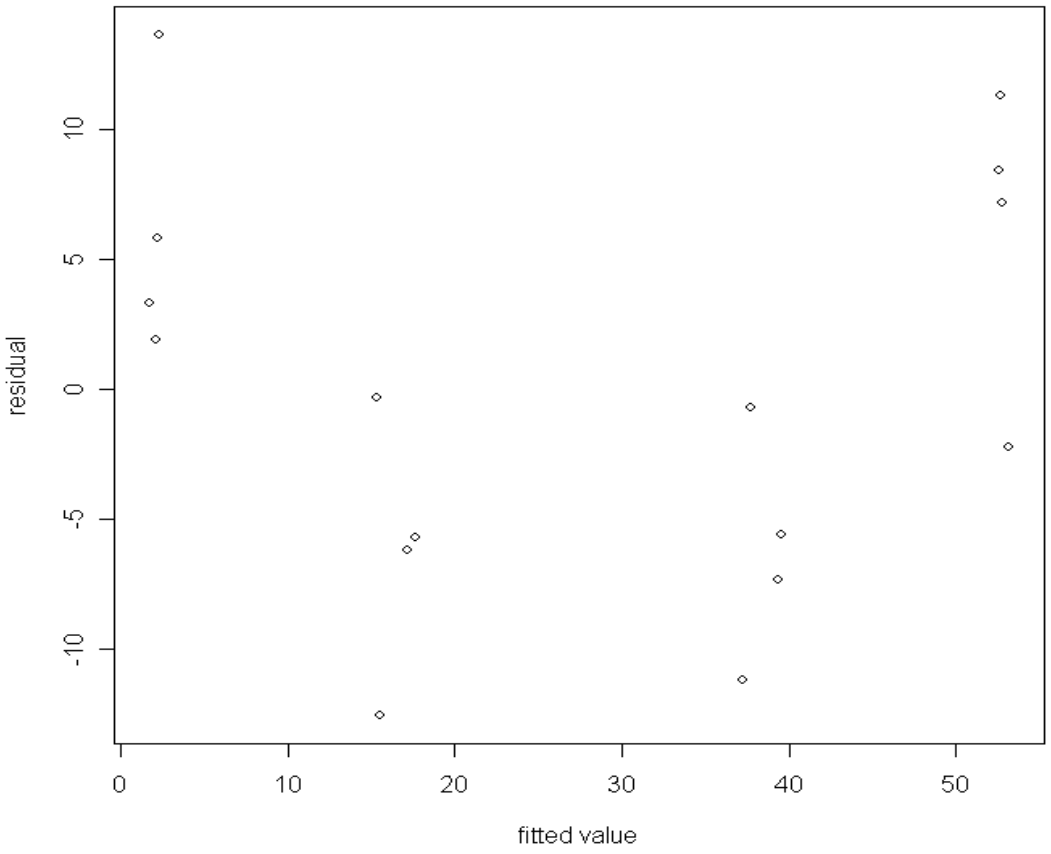
Question F2 is continued on the next page

(ii) The results of fitting a model that includes the main effects only, and a plot of the residuals against fitted values, follow.

```
Coefficients:

             Estimate   Std Error   t value     Pr(>|t|)
(Intercept)   27.4375      2.5414    10.796     1.88e-06
A              7.1875      2.5414     2.828       0.0198
B             18.1875      2.5414     7.156     5.33e-05
C             -0.4375      2.5414    -0.172       0.8671
D              0.1875      2.5414     0.074       0.9428
E              0.5625      2.5414     0.221       0.8298
F              0.0625      2.5414     0.025       0.9809

Estimated standard deviation of errors:  10.17
```



(a) How many degrees of freedom are available for the residual standard deviation? (1)

(b) Refer to the above plot and explain why you might have reservations about the adequacy of this model. (2)

Question F2 is continued on the next page

(iii) The results of fitting a model that includes the main effects and the maximum number of 2-factor interactions are shown below.

```
Coefficients:

             Estimate   Std Error   t value   Pr(>|t|)
(Intercept)   27.4375      1.9009    14.434    0.00477
A              7.1875      1.9009     3.781    0.06337
B             18.1875      1.9009     9.568    0.01075
C             -0.4375      1.9009    -0.230    0.83937
D              0.1875      1.9009     0.099    0.93042
E              0.5625      1.9009     0.296    0.79519
F              0.0625      1.9009     0.033    0.97676
A:B            6.1875      1.9009     3.255    0.08282
A:C           -0.9375      1.9009    -0.493    0.67071
A:D           -3.3125      1.9009    -1.743    0.22352
A:E           -0.6875      1.9009    -0.362    0.75223
A:F            0.0625      1.9009     0.033    0.97676
B:D           -0.3125      1.9009    -0.164    0.88453
B:F           -0.4375      1.9009    -0.230    0.83937

Estimated standard deviation of errors:  7.603
```

(a) Use the fitted model to recommend values for the control factors that are significant at the 20% level, and explain why you have made these recommendations.

(4)

(b) What reservations do you have about this model?

(1)

Question F2 is continued on the next page

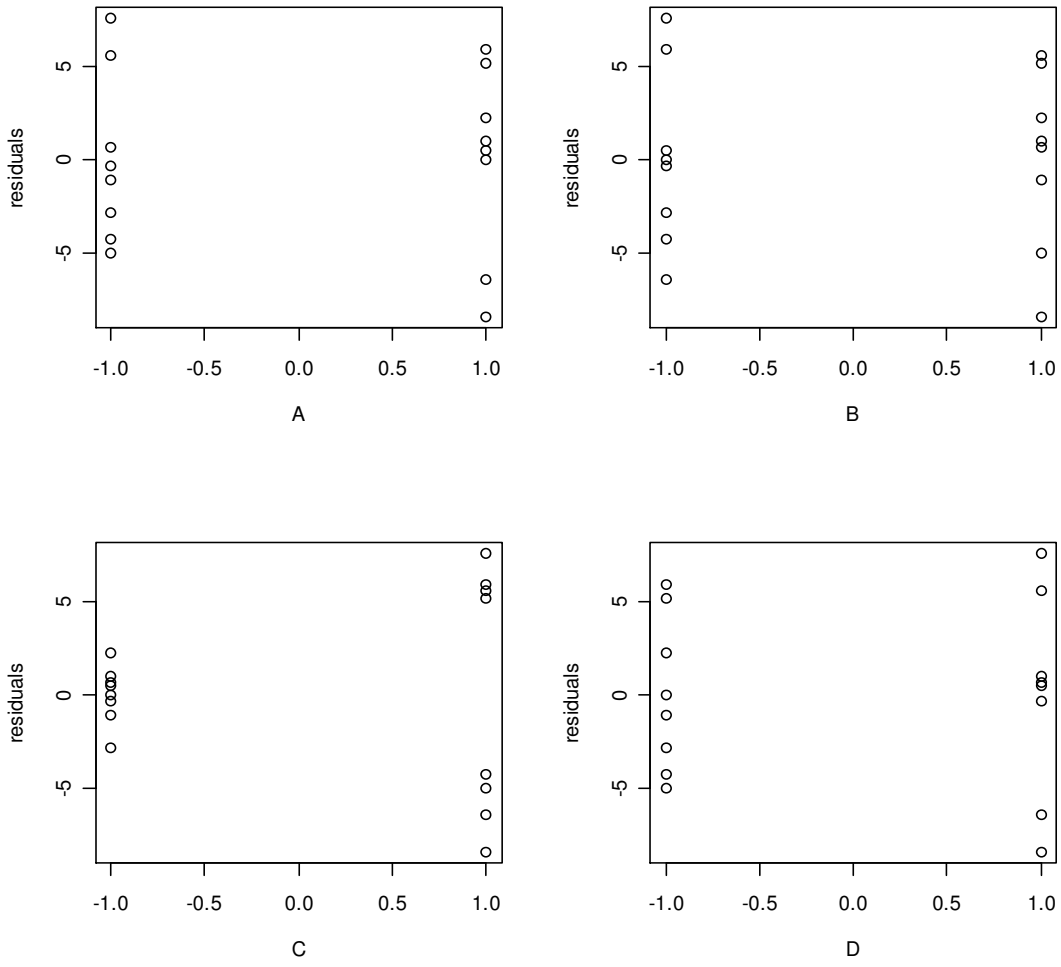
Turn over

(iv) Consider the following model and plots of residuals.

```
Coefficients:

             Estimate   Std Error   t value   Pr(>|t|)
(Intercept)   27.4375      1.5750    17.421   1.20e-07
A              7.1875      1.5750     4.564    0.00184
B             18.1875      1.5750    11.548   2.87e-06
C             -0.4375      1.5750    -0.278    0.78822
D              0.1875      1.5750     0.119    0.90817
E              0.5625      1.5750     0.357    0.73021
F              0.0625      1.5750     0.040    0.96932
A:B            6.1875      1.5750     3.929    0.00436

Estimated standard deviation of errors:  6.3
```



(a) Use the fitted model to recommend values for the control factors, and explain why you have made these recommendations**.** (4)

(b) The specification for porosity is that it must be less than 20. Make a rough estimate of the process capability using your recommended values from (a), and provide some justification for your estimate. (3)

Turn over

F3. The lifetime T of a machine component has cumulative distribution function (cdf) $F(t)$. A random sample of n such components has been drawn.

(i) Prove that the random variable $F(T)$ has a uniform distribution.

(3)

(ii) Let $T_{i:n}$ be the ith order statistic of the sample. Show that $E[F(T_{i:n})] = \dfrac{i}{n+1}$.

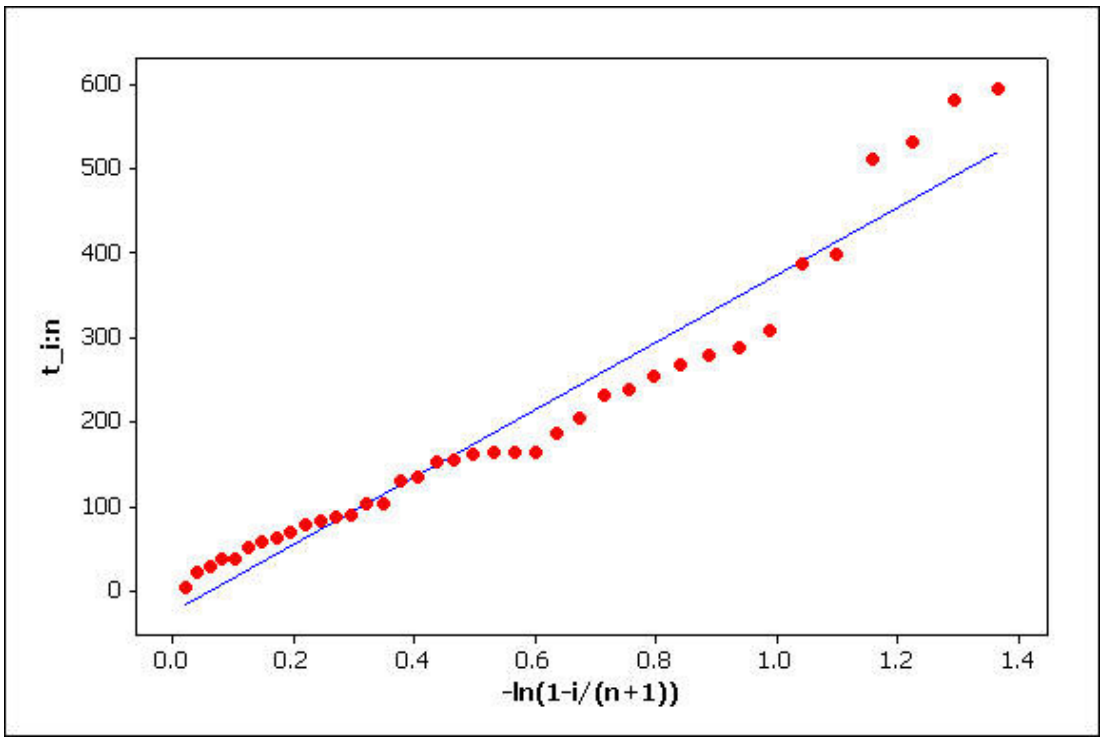
(7)

[You are given that if X_1, X_2, …, X_n is a random sample of size n from a probability distribution with cdf F and probability density function (pdf) f , then the pdf of the ith order statistic is

$$f_{X_{i:n}}(x) = \frac{n!}{(i-1)!(n-i)!} f(x)\{F(x)\}^{i-1}\{1-F(x)\}^{n-i} \ .\ ]$$

Question F3 is continued on the next page

(iii) A random sample of 50 pieces of mylar-polyurethane laminated DC HV insulating structure was selected. Each was tested at a voltage stress of 219 kV/mm for 10 hours. Thirty-eight failed and the mean survival time of these 38 pieces was 195.1 minutes. Twelve pieces survived the 600 minute test period. Suppose lifetimes have an exponential distribution with pdf $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, where $\lambda > 0$ is a parameter. The following is a plot of the ith smallest lifetime, $t_{i:n}$, against $-\log\left(1-\frac{i}{n+1}\right)$ for i from 1 to 38. [**Note.** "log", meaning logarithm to base e, is referred to as "ln" as the horizontal axis title on the plot.]



Given that the gradient of the fitted line is 397.2, estimate the rate parameter λ of the exponential distribution. Hence estimate the lifetime lower and upper first percentiles.

(3)

(iv) Derive a formula for the maximum likelihood estimator of λ, use your formula to estimate λ, and hence estimate the lifetime lower and upper first percentiles.

(5)

(v) What are the relative advantages of the two estimation methods?

(2)

Turn over

F4. (i) A machine has a lifetime that is exponentially distributed with a mean $1/\lambda$. The repair time has a distribution which is equivalent to the sum of two independent exponential distributions each with mean $1/(2\theta)$. The lifetimes and repair times are independent.

(a) What are the mean and variance of the repair time distribution? What is the general name of such a distribution?

(3)

(b) Set up a Markov chain model for the state of the machine.

(3)

(c) For what proportion of a long period of time is the machine operating?

(3)

(d) Comment briefly on the difference between this model and a model with an exponential distribution of repair time which has a mean of $1/\theta$.

(1)

(ii) A factory has two production lines, line 1 and line 2, for manufacturing car seats.

The preferred arrangement is to run both lines at standard speed. When running at standard speed the lifetime of line 1 has an exponential distribution with a mean of 30 days and the lifetime of line 2 has an exponential distribution with a mean of 15 days.

There is one repair crew and repair times are exponentially distributed with mean 2 days.

If one line fails the other can meet the production target if it is run at double speed. However, the means of the lifetime distributions are then reduced to 10 days and 5 days for lines 1 and 2 respectively.

If both lines fail the repair crew will repair line 1, because it is the more reliable, even if this means abandoning repair of line 2.

The lifetimes and repair times are independent.

(a) According to this protocol, if the repair crew is repairing line 2 and line 1 fails, the crew immediately moves to line 1, abandoning the repair on line 2 until later. Explain why it is unnecessary to allow for the time spent repairing line 2 in a model.

(1)

(b) Set up a Markov chain model for the state of the factory.

(4)

(c) Calculate the proportion of a long period of time that the factory is able to meet production.

(5)