EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA IN STATISTICS, 2009

Options Paper

Time Allowed: Three Hours

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

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

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

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

ANSWER EACH SECTION IN A SEPARATE ANSWER-BOOK.

Label each book clearly with its Section letter and title.

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

Graph paper and Official tables are provided.

Candidates may use 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* ${}^{n}C_{r}$.

This examination paper consists of 25 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 of Section A starts on page 2.
There are 24 questions altogether in the paper, 4 in each of the 6 Sections.

SECTION A – STATISTICS FOR ECONOMICS

A1. Briefly discuss issues associated with the selection of regressor variables in multiple linear regression analysis. Your account should include mention of redundant variables, best subsets regression, stepwise regression, residual analysis and model specification, and theoretical relationships.

(10)

Gujarati (2006) quotes data on the US annual demand for chicken (1960–1982). Summary information from four with-constant regression models is presented in the following table, the variables used being as follows.

Dependent variable Y: $\log_{10}$(per capita consumption of chicken (pounds))

X_1: $\log_{10}$(real disposable income per capita, in dollars)

X_2: $\log_{10}$ (real retail price of chicken per pound, in cents)

X_3: $\log_{10}$ (real retail price of pork per pound, in cents)

X_4: $\log_{10}$ (real retail price of beef per pound, in cents)

Model	R^2	Unusual observations	Durbin-Watson d	Regressor variables	Estd Coeffs (Std Errors)
A	0.982	None	1.83	X_1	0.343 (0.083)
				X_2	–0.505 (0.111)
				X_3	0.149 (0.100)
				X_4	0.091 (0.101)
B	0.982	1 std resid = 2.12	1.78	X_1	0.406 (0.045)
				X_2	–0.439 (0.083)
				X_3	0.107 (0.088)
C	0.980	None	1.87	X_1	0.441 (0.052)
				X_2	–0.382 (0.076)
				X_4	0.021 (0.092)
D	0.980	None	1.88	X_1	0.452 (0.025)
				X_2	–0.372 (0.063)

(i) For each regression model, specify the critical partial t value at the 5% significance level, assess each coefficient for partial significance and (if it is significant) discuss whether its sign is in accordance with economic theory.

(4)

(ii) Discuss the results of the analyses, stating with reasons which of models A, B, C and D is the most satisfactory in the light of all the information in the table above.

(4)

(iii) Suggest with brief justification what additional data or analysis would be helpful for the purpose of modelling Y in terms of X_1, X_2, X_3 and X_4.

(2)

A2. Purchasers of units in unit trusts hope to benefit from a combination of income paid by their trusts and growth in the value of their units. Equity Income Trusts specialise in high incomes, Growth Trusts in potential growth in the value of their units, while Growth and Income Trusts seek a compromise between these two objectives. Data have been collected to examine whether the three types of trust differed in their overall success over a recent 10-year period. Statistics of trusts' total returns, x, over the 10 years (i.e. the growths in their selling prices added to the incomes paid out, expressed as percentages of their selling prices at the start of the period) were compiled for samples of trusts of all three types specialising in investment within the UK, as follows.

UK Growth Trusts	121.0, 148.4, 225.0, 89.0, 135.2, 181.9, 169.6, 77.0, 150.9, 156.8 $\Sigma x = 1454.8$, $\Sigma x^2 = 228626.42$
UK Growth and Income Trusts	217.9, 183.1, 203.8, 195.8, 172.5, 245.6, 208.2, 223.1, 138.3, 199.7 $\Sigma x = 1988.0$, $\Sigma x^2 = 403081.54$
UK Equity Income Trusts	130.1, 171.5, 132.8, 123.9, 204.9, 216.8, 230.9, 225.9 $\Sigma x = 1436.8$, $\Sigma x^2 = 272657.18$

(i) Using analysis of variance, test the null hypothesis that there is no difference between the three types of trust as regards their mean total returns. State clearly the model, the standard assumptions and the null and alternative hypotheses for your analysis, and report your conclusion in terms that a non-statistician would understand.

(11)

(ii) Identify (but do not perform) a corresponding nonparametric procedure, stating clearly the null and alternative hypotheses and any necessary assumptions that have to be made about the data.

(3)

(iii) Which type of trust had the largest sample mean? For this type of trust, test the null hypothesis that the population mean was 210%. Also carry out a corresponding nonparametric test, stating the null and alternative hypotheses and explaining your method clearly.

(6)

A3. (i) One moving average trend which may be fitted to a series of observations $\ldots, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, \ldots$ is

$$y_t = \frac{x_{t-2} + 2(x_{t-1} + x_t + x_{t+1}) + x_{t+2}}{8} .$$

If the data x_t have been generated by a stationary process in which corr(x_t, $x_{t \pm i}$) = 0 for all t and all $i > 0$, find the autocorrelation function of the series y_t. (4)

(ii) Suppose that, instead of coming from a stationary process, the series x_t represents quarterly data in an economic time series. What advantage does the series y_t have over the series x_t in this instance? What further advantage does the series y_t have over the moving average trend series $z_t = (x_{t-1} + x_t + x_{t+1} + x_{t+2})/4$? (2)

(iii) A UK company's quarterly sales figures x_t (in thousands) for umbrellas for the years 2002–2007, and the corresponding values of y_t given by the formula in part (i) above, are shown in the following table.

Year	*Quarter*	*Sales* x_t	*MA* y_t	*Sales/MA*
2002	1	37.0	*	*
	2	39.5	*	*
	3	39.7	40.8125	0.97274
	4	45.4	41.5000	1.09398
2003	1	40.3	42.0000	0.95952
	2	41.7	42.2125	0.98786
	3	41.5	42.2625	0.98196
	4	45.3	41.9250	1.08050
2004	1	40.8	41.0375	0.99421
	2	38.5	40.0500	0.96130
	3	37.6	39.4250	0.95371
	4	41.3	39.1500	1.05492
2005	1	39.8	38.9875	1.02084
	2	37.3	38.5875	0.96663
	3	37.5	37.3375	1.00435
	4	38.2	36.0625	1.05927
2006	1	32.9	35.3750	0.93004
	2	34.0	35.0625	0.96970
	3	35.3	35.3375	0.99894
	4	37.9	35.5250	1.06685
2007	1	35.4	35.3125	1.00248
	2	33.0	35.1000	0.94017
	3	34.6	*	*
	4	36.9	*	*

Obtain multiplicative seasonal correction factors for these data, and use them to calculate seasonally-adjusted moving average trend values for the year 2006. (8)

(iv) Calculating additive seasonal effects as simple averages of the differences (sales − MA) for each quarter in turn is arithmetically simpler; why are multiplicative factors usually preferred? (3)

(v) Given that the quarters 1, 2, 3, 4 refer to (January – March), (April – June), (July – September), (October – December) respectively, interpret the estimated seasonal pattern in the data of this example. (3)

Turn over

A4. In order to examine the inter-relationship between male and female employment in the UK, seasonally-adjusted quarterly data are available giving numbers in thousands of males and of females in employment between Spring 2004 and Winter 2006 inclusive. Male numbers are denoted by M, female numbers by F, and a time trend $t = 1, 2, \ldots, 12$ is used. You are given that

$$\Sigma M = 185508 \qquad \Sigma F = 157177 \qquad \Sigma t = 78$$

and that the corrected sums of squares and products of M, F and t are

$$S_{MM} = 751734 \qquad S_{Mt} = 9201 \qquad S_{tt} = 143$$
$$S_{MF} = 600962 \qquad S_{Ft} = 8359 \qquad S_{FF} = 529123.$$

(i) Find the simple correlation coefficients r_{MF}, r_{Mt} and r_{Ft} and the partial correlation coefficient $r_{MF.t}$. Test the separate null hypotheses that each of these correlations is zero in the population from which these data were drawn as a random sample.

(7)

(ii) Find the OLS regression of M on t, and obtain the estimated standard errors of the coefficients.

(4)

(iii) The corresponding regression of F on t is

$$\hat{F} = 12718.1 + 58.455t$$
$$\qquad (39.2) \quad (5.322)\ ,$$

where the numbers in brackets represent estimated standard errors. The multiple regression of F on M and t, similarly expressed, is

$$\hat{F} = 6773.9 + 0.39521M + 33.026t$$
$$\qquad (1564.8)\ (0.10403) \quad (7.542)\ .$$

How do you explain the difference in the constant terms and in the coefficients of t in these two regression models for F?

How is the significance of the correlation coefficients (including the partial correlation coefficient) related to the significance of the coefficients in the regressions?

(4)

(iv) You are given that the multiple correlation coefficient $R^2_{F(Mt)} = 0.9706$ to 4 decimal places. Find the corresponding bias-adjusted value $\overline{R}^2_{F(Mt)}$.

(2)

(v) State with reasons which regression model for F you prefer, noting any reservations you may have.

(3)

SECTION B – ECONOMETRICS

B1. Consider the following structural two-equation model for $t = 1, 2, \ldots, n$.

$$y_{1t} + \gamma_{21} y_{2t} + \beta_{11} x_{1t} = \varepsilon_{1t} \quad [1]$$
$$y_{1t} + \gamma_{22} y_{2t} = \varepsilon_{2t} \quad [2]$$

The error terms $(\varepsilon_{1t}, \varepsilon_{2t})$ are not autocorrelated but may be contemporaneously correlated, (y_{1t}, y_{2t}) are endogenous and x_{1t} is exogenous. Also assume that $\gamma_{21} \neq \gamma_{22}$.

(i) Derive the reduced form equations.

(6)

(ii) Give necessary and sufficient conditions for equation [2] to be identified.

(4)

(iii) What is the importance of the assumption $\gamma_{21} \neq \gamma_{22}$?

(2)

(iv) Assuming that equation [2] is identified, describe how γ_{22} can be estimated by two-stage least squares.

(5)

(v) By following through the steps of two-stage least squares, show what will occur if you attempt to estimate the parameters of equation [1] by that method.

(3)

B2. (i) Suppose that a true econometric model for a series of annual data y_t is of the form

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \qquad [1]$$

where, for the years 1998, …, 2007,

$$x_{1t} = t = (\text{year} - 1997) = 1, \ldots, 10$$

$$x_{2t} = \begin{cases} 0 & t = 1, ..., 8 \\ t-8 & t = 9, 10 \end{cases}.$$

Throughout this question it is assumed that the $\{\varepsilon_t\}$ are identically and independently distributed N(0, σ^2) random variables.

(a) Interpret this model, making clear the role of the explanatory variables x_1 and x_2, and suggest an economic context for a model of this form.

(4)

(b) A researcher suggests that, due to a change of government policy late in 2001, the model should also provide for a constant step change in y_t for the years 2002, …, 2007. Define a further explanatory variable x_3 which would extend the model in this way.

(1)

(ii) Suppose that the correct model for y is

$$y_t = 1 + x_{1t} + 2x_{2t} + \varepsilon_t, \qquad [2]$$

so that in the notation of equation [1] above we have $\beta_0 = \beta_1 = 1$ and $\beta_2 = 2$. For the data for 1998–2007, relevant summary statistics are

$$\sum x_{1i} = 55, \quad \sum x_{1i}^2 = 385, \quad \sum x_{2i} = 3, \quad \sum x_{2i}^2 = 5, \quad \sum x_{1i} x_{2i} = 29.$$

(a) An econometrician omits the variable x_2 and fits instead the model

$$y_t = \beta_0 + \beta_1 x_{1t} + \varepsilon_t. \qquad [3]$$

Stating clearly any results assumed without proof, find the expected value of the OLS estimate of the parameter β_1.

Given that the estimated residual mean square from fitting model [3] is 0.9863, test on the basis of this model whether the parameter β_1 is significantly different from 1.

(8)

Question B2 is continued on the next page

(b) Suppose now that the correct model is of the form given in [3] with $\beta_0 = \beta_1 = 1$, i.e. $y_t = 1 + x_{1t} + \varepsilon_t$, but the econometrician fits the model given in [1], i.e. $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$. Without detailed calculation, explain with reasons how and why you would anticipate that the expectations and variances of the estimated coefficients in his fitted model would differ from those of the correct model.

(4)

(c) A student says that it is better to include an explanatory variable which is irrelevant that to risk omitting one which is relevant. Comment briefly on this statement.

(3)

B3. Based on a series of 26 annual observations from 1981 to 2006, the following model was estimated by least squares in an attempt to explain the volume of imports for a certain developing country.

$$\log \hat{y}_t = \underset{}{0.504} + \underset{(0.340)}{0.26 \log x_{1t}} - \underset{(0.109)}{0.108 \log x_{2t}} + \underset{(0.045)}{0.726 \log y_{t-1}} . \qquad \underset{(0.122)}{}$$

The figures in brackets below the parameter estimates are the associated estimated standard errors. In this equation,

y_t = volume of imports into the country (billion US$),

x_{1t} = real gross national product of the country (billion US$),

x_{2t} = index of import prices as % of an index of domestic prices in the country.

(i) Comment on the signs of the parameter estimates in this model, and test the constant and the exogenous terms in the model for partial significance. Is there a case for simplifying the model? Also test the null hypothesis of a unit coefficient for the lagged endogenous term.

(7)

(ii) Given that $x_{1\ 1995}$ = 33.784 billion US$, $x_{2\ 1995}$ = 156.7% and y_{1994} = 2.5193 billion US$, obtain a model-based estimate of y_{1995} in billions of US$.

(3)

(iii) Late in 1995 the country introduced reforms which either reduced or removed various barriers to free trade. Indicate ways in which the regression model could be modified to take account of the impact of these reforms.

(5)

(iv) Why is it important to test for autocorrelated errors in models such as this one? Outline a test for first-order autocorrelation.

(5)

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

(a) Misspecification.

(b) Multicollinearity.

(c) Cointegration.

(d) Instrumental variables.

(e) Heteroscedasticity.

SECTION C – OPERATIONAL RESEARCH

C1. (a) A sports shoe manufacturer makes four types of shoe: a running shoe, a cross-training shoe, a basketball shoe, and a terrain-running shoe.

To make a pair of shoes, a certain amount of time is required on each of a cutting machine (CM), a sewing machine (SM) and a finishing machine (FM). Each day there are 160 person hours available on the cutting machine, 178 person hours available on the sewing machine, and 144 person hours available on the finishing machine.

The profit from a single pair of shoes, and the amount of manufacturing time required on the different machines to make it, are given for each type of shoe in the table below.

	Profit (€)	*CM (hrs)*	*SM (hrs)*	*FM (hrs)*
Running shoe	120	0.5	1.0	1.0
Cross-training shoe	80	0.5	0.5	0.5
Basketball shoe	160	1.0	1.0	1.0
Terrain-running shoe	70	0.5	1.0	0.5

Formulate, but do not attempt to solve, a linear programming problem whose optimal solution stipulates how many pairs of each type of shoe should be made each day, in order to maximise the company's profit.

(6)

(b) (i) Use the Simplex algorithm to solve the following linear programming problem.

Maximise $z = 3x_1 + 5x_2$

Subject to

$$x_1 + 3x_2 \leq 9$$
$$2x_1 + 5x_2 \leq 12$$
$$x_1, x_2 \geq 0$$

(9)

(ii) Suppose that, at the first iteration of the Simplex algorithm above, you had chosen to pivot on the same column but the other row. Perform this pivot and hence write down the basic solution corresponding to the resulting tableau.

Giving your reasons, state whether or not this solution is feasible.

Using this example, justify the method used in the Simplex algorithm for selecting the pivot row.

(5)

C2. Consider the project management problem described by the following table:

Activity	*Duration (days)*	*Immediate predecessors*	*Personnel required*
A	8	—	4
B	3	—	6
C	4	B	10
D	4	A, C	4
E	2	B	10
F	8	E	3
G	1	D, F	6

(i) Draw the network representing this problem, compute the total float and free float for each job, and identify the critical path(s) and critical activities.

(9)

(ii) Sketch a Gantt chart for this project.

The project manager wishes to minimise the total number of people working on the project at any one time. Assuming that activities cannot be split once started, give a schedule for which the project can be completed in minimum time using no more than 17 workers.

(8)

(iii) Suppose that the durations of all activities are random variables, whose expected values are given in column two of the table above and whose standard deviations are as follows.

Activity	A	B	C	D	E	F	G
Standard Deviation	2	0	1	2	1	3	1

Making reasonable assumptions (which you should state clearly), find the approximate probability that the project will be completed in 20 days or less.

(3)

C3. Consider a retail company that regularly orders batches of stock from some manufacturer and then sells the items on to the public. In an Economic Order Quantity (EOQ) inventory model, demand is supposed constant at a rate of D per year, the ordering cost is £K for each order placed, the unit purchasing cost is £p per item, and the holding cost for one unit of inventory is £h per year. Suppose that there is no lead time between placing and receiving an order, and that no shortage of inventory is allowed.

(i) Describe what is meant by an *inventory cycle* and then derive the optimal size q_1 of an order (the EOQ) by minimising the annual cost.

(8)

(ii) Suppose now that instead of ordering stock from outside, the company produces it internally at a rate of r items per year ($r > D$). Setting up a production run costs £K, and each item costs £p to produce. Assume that there is no lead time between the setting up of a production run and the start of production, and that no shortage of inventory is allowed.

(a) Describe an inventory cycle for this continuous production model and hence obtain the optimal size q_2 of a production run. Show that it is always the case that $q_2 > q_1$.

(9)

(b) Suppose that the maximum level of inventory allowed is m. What is the optimal production policy under this restriction?

(3)

C4. Let $X(t)$ be the number of individuals at time t in an M/G/1 queuing system with arrival rate λ and service distribution given by the probability density function

$$f(x) = \begin{cases} 0 & x \le 0 \\ \dfrac{8}{11} & 0 < x \le 1 \\ \dfrac{8}{11x^3} & 1 < x \le 2 \\ 0 & x > 2 \end{cases}$$

Note that $X(t)$ includes any individual being served at time t.

(i) Give pseudo code for simulating from the density f using the inversion method.

(8)

(ii) Give pseudo code for simulating the number of arrivals during a time period of length T.

(6)

(iii) Let Y_n be the number in the system (those queuing and those currently being served) immediately following the nth departure, and let Z_n be the number of customers arriving during the nth service. Express Y_{n+1} in terms of Y_n and Z_{n+1}. Hence give pseudo code for simulating Y_{n+1} given Y_n.

(6)

SECTION D – MEDICAL STATISTICS

D1. (i) There are various randomisation methods for allocating patients to different treatments in a randomised controlled trial, including *block randomisation* and *stratified randomisation*. Briefly describe these two methods and contrast their use.

(6)

(ii) Compared with simple randomisation, how would the use of these two methods of randomisation affect the sample size of a trial?

(2)

(iii) A recent randomised controlled trial examined whether a new therapy was better than the standard therapy for reducing women's physical discomfort during routine gynaecological visits. Physical discomfort was measured on a 100 mm visual analogue scale and treated as a continuous variable in the analysis. The investigators wanted to ensure that they had 80% power to detect a 5 mm difference between the two treatment groups (SD for the treatment difference: 15 mm) at the two-sided 5% level. How many patients did they need in each group?

(7)

(iv) The trial was stopped after 204 patients had been enrolled, 100 in the standard therapy group, 104 in the new method. The investigators found a 13.2 mm difference in the visual analogue scale between the two methods in favour of the new method (95% confidence interval 6.8 to 19.7). Explain this treatment effect and its confidence interval in terms that would be understood by a non-statistician.

(2)

(v) Do you think that this study was underpowered? Give reasons for your answer.

(3)

D2. (i) As part of a study to examine potential differences between people with type 1 and type 2 diabetes, individuals with diabetes were monitored for six days to see if they experienced any episodes of hypoglycaemia. The time to first episode of hypoglycaemia was noted. Whilst time to event may be treated as a continuous variable and two groups compared using a *t* test, explain why this is not necessarily a good idea.

(3)

(ii) There were 50 patients in study group A and of these 39 had an episode of hypoglycaemia. In addition there were 57 patients in study group B and of these 46 had an episode. Using an appropriate statistical test, decide whether there was a statistically significant difference between the two groups in the proportion of patients who had an episode.

(6)

(iii) Below are the times (in hours) for the first 6 individuals in each study group. Perform a log rank test on these data. Is there a difference between the survival patterns of the two groups?

(11)

Regimen A: 53, 38, 123, 53, 130*, 24

Regimen B: 120*, 17, 21, 114, 134, 83

* indicates a censored observation

D3. (i) Explain what is meant by the term *prevalence* of a disease.

(2)

(ii) Explain what is meant by the term *incidence* of a disease.

(2)

(iii) The following data are taken from a recently published paper examining the association between self-rated health and mortality in different socioeconomic groups in the GAZEL cohort study. Calculate the relative risk of death during the follow-up period for men compared to women. Calculate the odds ratio of death during the follow-up period for men compared to women. Compare this odds ratio with the relative risk.

(3)

Table 1
Mortality at the end of the follow-up period
in the GAZEL study

	Men	*Women*	*Total*
Dead	902	157	1 059
Alive	13 964	5 357	19 321
Total	14 866	5 514	20 380

Question D3 is continued on the next page

(iv) These data were further broken down by occupational level. Using the Mantel-Haenszel procedure, calculate the occupation-level adjusted odds ratio of death during the follow-up period for men compared to women, after having adjusted for the confounding effect of occupational level.

(5)

Table 2(a)
Mortality at the end of the follow-up period in the GAZEL study: occupational level 1: unskilled

	Men	*Women*	*Total*
Dead	192	47	239
Alive	1 878	1 413	3 291
Total	2 070	1 460	3 530

Table 2(b)
Mortality at the end of the follow-up period in the GAZEL study: occupational level 2: skilled

	Men	*Women*	*Total*
Dead	521	98	619
Alive	7 755	3 496	11 251
Total	8 276	3 594	11 870

Table 2(c)
Mortality at the end of the follow-up period in the GAZEL study: occupational level 3: managerial

	Men	*Women*	*Total*
Dead	189	12	201
Alive	4 331	448	4 779
Total	4 520	460	4 980

(v) Is there evidence of a significant difference in the occupation-level adjusted odds of death during the follow-up period for men compared to women? Comment on the results of this hypothesis test.

(5)

(vi) Calculate the 95% confidence interval for the adjusted odds ratio in (iv). Comment on the odds ratio and its 95% confidence interval.

(3)

D4. (i) What distinguishes a *screening test* from a *diagnostic test*?

(4)

(ii) Define the *sensitivity*, *specificity* and *positive predictive value* of a diagnostic test.

(3)

(iii) Researchers were interested in whether hypoglycaemia (as indicated by a blood glucose reading of less than 3.0 mmol/l) could be detected using a new continuous glucose monitoring system (CGMS). They studied 223 people with diabetes, of whom 117 had a blood glucose reading of less than 3.0 mmol/l. They investigated the usefulness of three different cut-offs for CGMS: 3.5, 3 and 2.2. The results are presented below. Estimate the corresponding test sensitivities and specificities for each of these cut-offs and sketch the ROC curve for the CGMS technology.

(10)

Cut-off	*CGMS classification*	*Blood Glucose*	
		Hypoglycaemic	*Not hypoglycaemic*
3.5	*Hypoglycaemic*	105	50
	Not hypoglycaemic	12	56
3.0	*Hypoglycaemic*	82	36
	Not hypoglycaemic	35	70
2.2	*Hypoglycaemic*	49	12
	Not hypoglycaemic	68	94

(iv) What would be the shape of the ROC curve for a perfect diagnostic test? Does the ROC curve obtained in (iii) suggest that this test would be any use in practice?

(3)

SECTION E – BIOMETRY

E1. Describe carefully two ways in which a *negative binomial* distribution may arise.

State two one-parameter distributions which are special cases of the negative binomial family.

(5)

A random sample of 100 stalks taken from a field of maize infested with corn-borers gave the following distribution of corn-borers per stalk.

Number of borers, r	0	1	2	3	4	5	6 or more
Number of stalks, f	23	12	26	24	11	4	0

Investigate the null hypothesis that the borers are distributed at random

(a) by comparing the variance and mean of the sample,

(3)

(b) by carrying out a χ^2 goodness-of-fit test.

(8)

What features of this distribution might lead to contradictory conclusions from these two methods?

(4)

E2. Explain briefly the importance of *fractional replication* and *confounding* in experimental design.

(4)

An experimenter is investigating the response of a corn crop to changes in the applied levels of six fertiliser elements A – F. The available site divides naturally into 2 blocks, each containing 18 plots. He therefore decides to use most of these plots to carry out a half-replicate of a 2^6 factorial. Give full details of a design for this experiment, and list the items in its analysis of variance, with their degrees of freedom. (Formulae for sums of squares are not required.)

(14)

The experimenter is concerned that the response curves for two of these elements (C and D) might possibly be non-linear. Suggest how he could use the remaining available plots to examine this possibility.

(2)

E3. A report reads as follows.

"In this investigation into the effect of diet on blood sugar, it was decided to compare 3 diets and 4 strains of rat. Each diet was allotted to 4 rats, one chosen at random from each strain. The diets were fed for 4 weeks, and at the end of each week duplicate determinations of blood sugar were made on each rat. The Analysis of Variance of the results was as follows.

	D.F.	Sum of Squares	Mean Square	*F*
Replication	1	0.19		1.73
Diets	2	6.03	3.02	27.45
Strains	3	5.82	1.94	17.64
Diets × Strains	6	3.18	0.53	4.82
Weeks	3	7.47	2.49	22.64
Diets × Weeks	6	6.12	1.02	9.27
Strains × Weeks	9	6.51	0.72	6.55
Diets × Strains × Weeks	18	10.02	0.56	5.09
Error (residual)	47	5.17	0.11	
TOTAL	95	50.51		

The values of *F* for diets, strains, weeks and all their interactions are significant at the 0.1% level. We were surprised to find such very significant interactions between weeks and diets, and weeks and strains, and assumed at first that this was caused by some diets and some strains reaching saturation earlier than others; however, an investigation of the means shows that this was not so. An explanation of these interactions must therefore await further study."

Illustrate the design of the experiment by a sketch, and comment on it.

(5)

In the light of your sketch, comment critically on the analysis and conclusions contained in the report. Provide an appropriate analysis, and state your conclusions clearly.

(15)

E4. Define the terms *tolerance* and *ED50* as used in bioassay.

(4)

To assess the tolerance of snails to a certain pesticide, six different concentrations were applied to six randomly chosen groups of snails. Numbers of snails in each group, and numbers that died in each group, are shown in the table.

Concentration (mg/l)	10.7	8.2	5.4	4.2	2.3	0
Number of snails	60	60	57	60	60	59
Number of snails dying	53	52	30	20	7	0

Plot two graphs, one using concentration on the horizontal axis and one using log(concentration), which could be used in estimating the ED50 of tolerance in the snail population.

(4)

What assumptions about the tolerance distribution are involved in choosing between a linear model using concentration and one using log(concentration) as the explanatory variable?

(2)

Estimate the ED50 of the pesticide.

(2)

Outline a method of analysis of these data that could be carried out by computer to give an estimate of ED50 and a confidence interval for it.

(8)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A company (M) that manufactures electronic calculators buys the casings from a supplier (S). There is a detailed specification for the casings that includes the quality requirements of the hinges, dimensions, and surface finish.

(i) M and S agree to the following acceptance sampling procedure. The casings will be delivered in batches of 1000. M will inspect a random sample of 30 casings. A casing will be classed as defective if it does not meet all aspects of the specification. The batch will be accepted if 0 or 1 defectives are found. If more than one defective is found the entire batch will be inspected, and S will pay for this inspection and replace all defective items.

(a) Calculate the probabilities of accepting a batch, without full inspection, if the proportion of defectives (p) is: 0.04; 0.05; 0.06. If you make an approximation, justify its use. (3)

(b) Calculate the average outgoing quality (AOQ) of the acceptance sampling procedure, if the proportion of defectives is: 0.04; 0.05; 0.06. (2)

(c) Define the average outgoing quality limit (AOQL) for the acceptance sampling procedure. Without making any further calculations, sketch the graph of AOQ against p and give an approximate value for the AOQL. (2)

(d) Calculate the producer's risk if the proportion of defectives in the batch is 0.02. (2)

(e) Discuss, briefly, the benefits and drawbacks of the acceptance sampling procedure, and the advantages and disadvantages of relying on a single supplier. (4)

(ii) S has set an internal specification for casings which is that 80% do not break if they are dropped on to a concrete floor from a height of 2 m. A random sample of 30 casings has been taken each week. The numbers of casings that have broken over the past four weeks are 4, 7, 8 and 14.

(a) Set up a p-chart for the proportion of casings that fail, showing the upper action line, and plot the results from the last four weeks. Comment on your chart. (5)

(b) Explain why, in general, a lower action line is useful on p-charts when plotting proportions of samples that fail. (1)

(c) Why might S choose to plot proportion of casings that fail rather than the number of casings that fail? (1)

F2. A company manufactures circuit breakers for high voltage power applications. A critical characteristic of the circuit breakers is the open time, and this is measured for every circuit breaker produced. Denote the sequence of open times, which are recorded in tenths of micro-seconds, by $\{Y_t\}$, and assume open times are independently distributed.

The target value for the open time is τ and the standard deviation of open times is σ. An engineer intends to set up an exponentially weighted moving average (EWMA) chart defined by:

$$\tilde{Y}_1 = \theta Y_1 + (1-\theta)\tau$$
$$\tilde{Y}_t = \theta Y_t + (1-\theta)\tilde{Y}_{t-1} \quad \text{for } t = 2, 3, \ldots .$$

(i) Suppose the process mean is equal to the target value, τ. Show that $E[\tilde{Y}_t] = \tau$ for all t, and derive the standard deviations of $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$ in terms of θ and σ. Assume that the variance of $\tilde{Y}_t$ tends to a constant value as t becomes large, and hence, or otherwise, show that the standard deviation of $\tilde{Y}_t$ is approximately $\sqrt{\frac{\theta\sigma^2}{2-\theta}}$ for large values of t.

(8)

Suppose further that $\tau = 100$, $\sigma = 4$ and $\theta = 0.2$.

(ii) Calculate the upper action limit for the EMWA for large values of t.

(2)

(iii) The specification for open time is that it should be within the range [85, 115]. Calculate the process capability index. Calculate the process performance index if the mean of the process is 105.

(2)

(iv) Now suppose the mean of the process jumped to 105 a few time steps ago. The latest value of the EWMA is 103.4. What is the probability that action will be indicated at the next time step?

(3)

(v) Consider the following algorithm. Let the target value be τ, the standard deviation of the open times be σ, and $K = 0.5\sigma$. Set $SH(0) = 0$, $SL(0) = 0$, and define

$$SH(t) = \max[0, (y_t - \tau) - K + SH(t-1)]$$
$$SL(t) = \max[0, -(y_t - \tau) - K + SL(t-1)]$$
$$\text{for } t = 1, 2, \ldots .$$

Action is indicated if either $SH(t)$ or $SL(t)$ exceeds 5σ. Calculate $SH(t), SL(t)$ for t from 1 to 5 if the first five open times are 108, 103, 94, 111 and 114. State with a reason whether action is indicated.

(5)

Turn over

F3. Twelve disc drives for notebook computers were subjected to a highly accelerated lifetime test of 360 hours duration. Seven failed after 71, 147, 189, 197, 216, 312 and 353 hours respectively. Two were randomly selected from amongst those still working at 240 hours, and withdrawn for detailed inspection. The other three lasted longer than 360 hours. Let T denote the lifetime of a randomly selected disc drive under these conditions. It is thought that T has the Weibull distribution with cumulative distribution function

$$F(t) = 1 - \exp(-\lambda t^{\alpha}) \qquad (t > 0),$$

where $\lambda > 0$ and $\alpha > 0$ are parameters to be estimated.

(i) Show that $\log\big(-\log(1 - F(t))\big) = \log(\lambda) + \alpha \log(t)$.

(2)

(ii) Use the result in (i) to construct a plot to check the Weibull assumption. Give your conclusion and use this plot to obtain estimates of the two parameters by eye.

(13)

(iii) Use your estimates obtained in (ii) to estimate the probability that a randomly selected disc drive subjected to these conditions that has survived 360 hours will survive a further 24 hours.

(2)

(iv) Write down the likelihood function for the data based on the Weibull distribution. Explain how you would find the maximum likelihood estimates of the parameters, given your likelihood function, but do not attempt any numerical calculations.

(3)

F4. (a) Let $\psi_S(x_1, \ldots, x_n)$ and $\psi_P(x_1, \ldots, x_n)$ represent the structure functions for n components in series and in parallel respectively. The $x_1, \ldots, x_n$ represent the states of the components, and take the values 1 or 0 for a working or failed component respectively.

(i) Sketch a block diagram for a system having the structure function

$$\psi_S(\psi_P(x_1, x_2),\ \psi_P(x_3, x_4, x_5),\ \psi_P(x_6, x_7)).$$

Write down the structure function explicitly in terms of $x_1, \ldots, x_7$.

(2)

(ii) Sketch a block diagram for a system having the structure function

$$\psi_P(\psi_S(x_1, x_3, x_6),\ \psi_S(x_2, \psi_P(x_4, x_5), x_7)).$$

Write down the structure function explicitly in terms of $x_1, \ldots, x_7$.

(2)

(iii) Use the block diagrams to explain why the system given in (ii) is less reliable than the system given in (i). Calculate the reliabilities of the two systems if the components fail independently and each has reliability 0.8.

(5)

(iv) Now suppose further that the probability that x_4 fails given that x_5 has failed and the probability that x_5 fails given that x_4 has failed are both 0.5. Calculate the reliability of the system given in (ii).

(4)

(b) A factory has two identical machines that operate independently. The times between failures for each machine are exponentially distributed with mean 4 hours.

There are two repair men who individually repair machines at a mean rate of 0.6 repairs per hour. The men do not cooperate on repairs. The repair times are independently exponentially distributed.

(i) For what proportion of time are 0, 1 and 2 machines working?

(5)

(ii) What is the mean number of machines in operation?

(1)

(iii) What is the mean number of man hours per 8 hour day during which the repair men are idle?

(1)