EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2010

MODULE 6 : Further applications of statistics

Time allowed: One and a half hours

Candidates should answer **THREE** *questions.*

Each question carries 20 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 5 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. Data were collected on the output from an industrial process carried out either by a standard method S or by one of four alternative methods A, B, C, D. There were 25 experimental units available, the order of experimentation was completely randomised, and the methods had different numbers of replicates. The following table gives the results y in suitable units.

	Total	*Mean*
S: 24.2, 27.7, 21.8, 30.6, 24.9, 29.3, 28.4, 26.7	213.6	26.7
A: 31.8, 37.7, 32.9, 40.2, 42.4	185.0	37.0
B: 35.7, 40.9, 50.4, 43.8	170.8	42.7
C: 40.2, 39.5, 46.3, 44.8	170.8	42.7
D: 30.6, 33.3, 36.2, 37.5	137.6	34.4
	877.8	

$\Sigma y^2 = 32\,186.08$

(i) Construct an analysis of variance table for these data, and test whether there is evidence of differences among the mean outputs from the five methods.

(7)

Write down the linear model which is the basis for this analysis, and state the assumptions that have to be satisfied by the terms in the model.

(2)

(ii) Construct a 95% confidence interval for the difference between the mean outputs from the standard method S and the method A.

(3)

(iii) In a completely randomised experiment, the residual for a unit is the difference between the value observed and the mean of all units that received the same treatment. For example, the mean for S is 26.7 and so the residuals for the units receiving S are –2.5, 1.0, –4.9, 3.9, –1.8, 2.6, 1.7, 0.0.

Calculate the residuals for A, B, C, D.

(3)

(iv) Draw a dotplot of all the residuals, showing the five methods on separate rows. Use this plot to comment on the assumptions that you have stated in your answer to part (i).

(5)

Turn over

2. (i) A set of n triples of observations (x_i, y_i, z_i) of variables X, Y, Z is available (i = 1, 2, 3, …, n) and the model $Y = \beta X + \gamma Z$ (without a constant term) is to be fitted to them.

Derive the least-squares estimators of β and γ.

(10)

(ii) (a) Show how the results of part (i) can be used to fit a quadratic relation $Y = \beta X + \gamma X^2$ between two of the variables.

(4)

(b) Suppose now that there is some doubt whether the quadratic relation in part (ii)(a) should go through the origin or not. Suppose also that a computer program for multiple regression is available. Explain how you would use this program to compare the fit of the model in part (ii)(a) and a model which also contains a constant term α. Suggest suitable diagnostics to use in this comparison.

(6)

Turn over

3. In an agricultural experiment, three levels (L1, L2, L3) of a nutrient were applied to a crop at each of four times (T1, T2, T3, T4) during the growing season. The experiment was laid out in three randomised blocks, giving altogether 36 unit plots.

The table below is a summary of crop yields, *y*, recorded at the end of the season. The table gives the **totals**, over the three blocks, of each level-time combination.

Total yield from three blocks

		Level			
		L1	L2	L3	*Time total*
Time	T1	42	54	70	166
	T2	51	62	83	196
	T3	76	84	98	258
	T4	76	89	106	271
Level total		245	289	357	891

The sum of the squares of all 36 individual records was $\Sigma y^2 = 23\,649$. Block totals were 265, 294 and 332.

(i) Draw an appropriate graph showing all twelve means (or totals) of T, L combinations.

(4)

(ii) Copy and complete the following analysis of variance.

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	*F* VALUE
Blocks	2	***	***	***
T	***	***	***	***
L	***	***	***	***
T × L	***	***	***	***
Treatments	***	1375.42		
Residual	***	***	***	
TOTAL	35	1596.75		

(10)

(iii) Write a report on the results of your analysis, carrying out any necessary statistical tests to justify your conclusions.

(6)

4. In a routine process for checking the weight of a packaged food product, two possible schemes are proposed for examining each large batch of the product. Experience is that 10% of the packaged items are outside specified tolerance limits ("faulty").

Scheme I takes a single sample of $n = 24$ items, and rejects the batch if three or more items are faulty.

Scheme II is a double sampling scheme, using samples of $n = 12$ items. The whole batch is accepted if there are no faulty items in the first sample, and is rejected if three or more are faulty. When there are one or two faulty items in the first sample, a second sample of 12 items is taken and the combined number of faulty items in both samples is considered. If this number is not greater than two, the batch is accepted and if it is three or more the batch is rejected.

(i) Find the probability of rejection using Scheme I.

(4)

(ii) For Scheme II find the probability of

(a) accepting the batch after taking only the first sample,

(1)

(b) rejecting the batch after taking only the first sample,

(4)

(c) rejecting the batch after taking both samples.

(4)

(iii) Find the overall probability of rejection of a batch using Scheme II and compare it with that using Scheme I.

(2)

(iv) Using the results in part (ii), find the probability of needing the second sample using Scheme II, and hence find the average sample size for Scheme II.

Comment on which scheme is likely to be preferred, on the basis of sample size and probability of rejection.

(5)