EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2008

(Modular format)

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* ${}^{n}C_{r}$.

This examination paper consists of 4 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. In an apple storage experiment, five batches from each of five groups (A to E) of apples were stored at a constant temperature under constant conditions. The experiment was laid out as a Latin square; the storage chamber had five rows of apples at different vertical heights, and the apples were arranged in five columns at different distances from the door. After storage, the percentage weight loss was recorded for each of the 25 experimental units (batches). The following table shows the layout of the experiment and the percentage weight loss during storage for each batch.

		Column					
		1	2	3	4	5	*Total*
Row	1	C 18.30	B 35.25	D 30.32	A 16.08	E 42.85	142.80
	2	D 28.05	E 36.16	A 17.25	C 25.90	B 31.98	139.34
	3	A 25.12	D 28.55	B 37.10	E 38.27	C 23.68	152.72
	4	B 40.25	C 22.60	E 41.15	D 31.68	A 22.15	157.83
	5	E 34.24	A 26.42	C 15.05	B 36.52	D 33.20	145.43
	Total	145.96	148.98	140.87	148.45	153.86	738.12

Totals for treatments were:

A 107.02; B 181.10; C 105.53; D 151.80; E 192.67.

(i) Copy and complete the following Analysis of Variance table.

(9)

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	*F* value
Rows	***	45.3682	***	***
Columns	***	17.9588	***	***
Treatments	***	***	***	***
Residual	***	***	***	
TOTAL	***	1583.1384		

(ii) Groups A and B contained the same variety of apple but stored for different lengths of time. Construct a 95% confidence interval for the difference between the means of all apples in groups A and B, and state carefully the inferences that can be made from this interval.

(4)

(iii) Groups C, D and E contained another variety; C was stored for a shorter time than D and E, which had equal storage times. Carry out a significance test of the null hypothesis that there was no effect of storage time.

(4)

(iv) Group D had a protective covering while in store; group E did not. Carry out a significance test of the null hypothesis that there was no effect of covering.

(3)

2. (a) A regression model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ is to be fitted to a set of data consisting of the n triples (y_i, x_{1i}, x_{2i}), $i = 1, 2, \ldots, n$.

Derive the normal equations for obtaining least squares estimates of α, β_1, β_2.

(7)

(b) The plasma lipid level of total cholesterol (y) was recorded for a sample of 8 patients before they received drug therapy. The investigator wanted to see how, if at all, these levels depend on the weight (x_1) and the age (x_2) of the patient.

From analyses of the data, the following regression sums of squares were obtained.

Regression on weight alone	0.1486
Regression on age alone	4.0439
Regression on weight and age	4.1139.

The total sum of squares about the mean was 4.9150.

Using backwards elimination or otherwise, obtain the best model for describing the relationship between Y and the x-variables.

(13)

3. (a) (i) Give an example where a (dummy) indicator variable would be useful in a regression analysis of a set of data. Write down the model that would form the basis of the analysis and explain the meaning of each of the parameters in it.

(6)

(ii) Give an example where a quadratic regression through the origin would be a suitable model to fit to a set of data (y_i, x_i), $i = 1, 2, \ldots, n$. Explain how the estimated regression equation could be used to estimate the maximum value of y. How would you report on the result if this estimated maximum occurred for a value of x that was outside the range of the available data?

(6)

(b) Explain briefly the uses of residuals in deciding whether a simple linear regression is an adequate explanation of a set of data. Illustrate your explanation with diagrams that are commonly available in computer package output and with statements that may appear in the output.

(8)

4. A continuous production process manufactures resistors which are intended to operate in normal use at 40μΩ. Tests on 20 random samples, each of 3 resistors chosen from batches as they came off the production line, gave the data summarised in the following table. From past experience it can be assumed that the standard deviation of a measurement is 3.2μΩ.

Batch	1	2	3	4	5	6	7	8	9	10
Total of 3 measurements in sample	117	121	110	120	114	120	127	120	126	125
Range of 3 measurements	3	3	4	8	2	6	5	5	6	4

Batch	11	12	13	14	15	16	17	18	19	20
Total of 3 measurements in sample	125	135	130	126	131	136	135	143	143	148
Range of 3 measurements	7	3	7	7	9	6	6	3	6	7

Construct two control charts and use them to discuss whether you might have concluded that the process was out of control at any stage as the batch results came in.
(12, 8)

[You are given that for samples of size 3 from a Normal distribution, the factors 0.04, 0.18, 2.17, 2.99 when multiplied by the average range will give the warning and action limits in a control chart of the range.]