EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2012

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 $\genfrac{}{}{0pt}{}{n}{r}$ *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 a food production process, packaged items are sampled as they come off a production line. A random sample of 5 items from each large production batch is checked to see whether each item is tightly packed. A batch will be accepted immediately if all of these 5 items are tightly packed, and rejected immediately if at least 3 items among the 5 are not tightly packed. Otherwise a second sample is taken before making a decision.

Suppose that 80% of the items produced by the machine are tightly packed.

(i) What is the probability that the second sample will be needed?

(5)

(ii) When a second sample is to be taken, it also consists of 5 items. What is the probability that in a combined sample of 10 items there are at least 8 tightly packed?

(7)

(iii) Suppose that a batch will be accepted if there are either 5 tightly packed items in the first sample or a total of at least 8 tightly packed items in the two samples together (taken according to the rules given above). Find the total probability of accepting a batch.

(4)

(iv) Bearing in mind that batches are large, and all the items in rejected batches must be inspected, explain briefly whether you consider that this is a good scheme or not. Suggest any changes that might improve it.

(4)

2. Two factors, A and B, were included in an experiment that was laid out in a randomised block design containing three blocks. Three levels a_0, a_1, a_2 of A were used, and two levels b_0, b_1 of B. The table below gives the totals of the observations in the three blocks for each of the six treatment combinations a_0b_0, …, a_2b_1.

	Level of A			
	a_0	a_1	a_2	*Total for B*
Level b_0	56	78	97	231
Level b_1	68	75	78	221
Total for A	124	153	175	452

The totals for the six observations in each block are:

block I, 136; block II, 151; block III, 165.

The total (uncorrected) sum of the squares for all 18 observations is 11 772.

Construct the analysis of variance for these data, dividing the treatments sum of squares into components for each main effect and the interaction. What inferences can be drawn from this analysis?

(13)

Draw a graph, showing the totals given in the table above on the vertical axis and the levels (equally spaced) of A on the horizontal axis, and use different symbols for the two levels of B. Use the graph and the analysis to explain more fully the results of the experiment.

(7)

3. (i) On a set of n experimental units, triplets of observations (x_i, y_i, z_i), i = 1 to n, are obtained. A scientist considers that the linear model

$$y_i = \alpha x_i + \beta z_i + \varepsilon_i$$

might fit these data.

Derive the ordinary least-squares estimates of the parameters α and β, and explain what α and β represent.

(12)

Suppose now that the scientist wants to replace z by the product xw in the model above, using a set of data (x_i, y_i, w_i). Write down the estimates of α and β using this new model.

(3)

(ii) Someone asks whether the original model ought to have a constant term (μ) added to it. If a suitable computer program for multiple regression is available, suggest two ways in which the scientist can compare the two versions, the original model given in part (i) above and the model including a constant term.

(5)

Turn over

4. (a) A biologist is studying how two different compounds affect the growth of plants. The compounds are used at similar levels x and measurements of growth y are taken after a fixed period of time. Draw sketch graphs that illustrate each of the following growth situations, and write down the equations that represent them.

(i) Two parallel straight lines.

(ii) Two straight lines having the same intercept but different gradients.

Show how each of (i) and (ii) can be written in a single equation using an indicator variable. Explain what information the indicator variable gives in each case.

(8)

(b) (i) Explain the meaning of the word *linear* in the phrase "the linear model".

Show that when $y > 0$ and $x > 0$, the relation $y = a\exp(kx)$, in which a and k are constants to be estimated, can be expressed as a linear model. If the intercept in the linearised model has the point estimate 3.51, with 95% confidence interval (2.65, 4.37), calculate the point estimate for a, and the corresponding confidence interval.

Suppose that someone sees your answer and says it must be wrong because the point estimate is not in the centre of the interval. What would be your response?

(6)

(ii) The three models given below may be used to explain the dependence of a random variable y on x, or on x_1 and x_2, where a, b, c, k are constants to be estimated.

For each of these models, state whether it is linear. For any models that are not linear, give a brief reason why. Give the linearised version when there is one, and explain briefly how the parameters can be estimated.

$$y = ax_1^b x_2^c$$

$$y = a + b(1 - e^{-cx})$$

$$y = ax^b e^{-kx}$$

(6)