EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2016

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 4 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. A company is considering three different online learning packages for training its staff to perform a certain set of tasks, and runs a small trial based in five of their offices in different cities. They know that offices vary in their overall skill at these tasks, but are confident that the learning packages are similar enough that they will not have different effects in different offices. In each office, one subject is allocated to each of packages A, B and C. Each subject independently completes the allocated learning package and then takes an online test.

(i) Explain the concept of *randomisation*, using this experiment as an example. Comment on the importance or otherwise of randomisation in this experiment.

(4)

(ii) Explain the concept of *blocking*, using this experiment as an example. Comment on the importance or otherwise of blocking in this design and on additional forms of blocking that might be helpful.

(5)

(iii) The test is the same for all subjects and is marked out of 100. The results are as follows. Write down an appropriate analysis of variance model for these data, stating and justifying any assumptions made.

(3)

	Office					
Package	1	2	3	4	5	Sum
A	77	49	78	64	72	340
B	28	75	38	63	29	233
C	51	59	49	54	51	264
Sum	156	183	165	181	152	837

$$\Sigma y^2 = 50\,357$$

(iv) Construct the analysis of variance table for these results. State appropriate hypotheses and test at the 5% level whether there is any evidence of a difference in mean test score between the learning packages. Clearly state your conclusions.

(8)

2. (a) A trial is to be conducted to investigate the effectiveness, for obese people trying to lose weight, of different diets and of different intervals between consultations with a dietary specialist. The trial involves four different diets and four different intervals (1, 7, 14 and 28 days), and all of the trial subjects are females aged 30–50 years.

Explain the following terms using this experiment as an example, outlining how the trial would be designed and why these concepts are important both in general and in this case. In particular, you should mention the advantages of applying these ideas and outline the reasons for these advantages. You should mention the analysis of the results where this helps to explain reasons for the design, but you should not go into the details of the analysis.

(i) Factorial designs.

(5)

(ii) Interactions and interaction terms.

(3)

(iii) Blinding.

(2)

(b) Define the *residuals* from a fitted multiple regression model. Hence explain how the residuals, and in particular residual plots, can be used to check each of the assumptions made by the model. Use sketch diagrams to illustrate your discussion where appropriate.

(10)

3. An industrial process is defined to be in control when the masses of the items it produces are Normally distributed with mean 132 g and standard deviation 4 g. At regular intervals a batch of five components is removed from the production line and their masses measured.

Successive batches produce mean masses (in g) as follows.

132.87, 133.52, 136.29, 135.41, 136.44, 135.17, 138.01, 136.55, 137.86.

(i) Explain how to construct a Shewhart control chart for the mean mass. Define and calculate 95% warning limits and 99.9% action limits. Explain and justify the ideas and assumptions underlying this control chart.

(6)

(ii) Use the Shewhart control chart to decide whether this process is out of control.

(3)

(iii) For a process defined as being in control in terms of a target value for the mean, explain how a cusum chart would be set up, how it works and why it is useful, indicating how it would look both if the process were in control and if it were not. Compare the usefulness of a cusum chart with that of a corresponding Shewhart chart.

(7)

(iv) Construct and interpret a cusum chart for the process and data above.

(4)

4. The multiple linear regression model

$$Y = \alpha + \beta w + \gamma x + \varepsilon$$

is fitted to n data points (w_i, x_i, y_i), $i = 1, 2, \ldots, n$.

(i) Explain what the random variable ε represents and the assumptions about it required by this model.

(2)

(ii) Derive the normal equations for finding the least squares estimators for α, β and γ for this model.

(7)

(iii) Explain the concept of *indicator variables* and hence show how this multiple linear regression model can be used as a one-way analysis of variance model with three groups, with n_j observations in the jth group, where the third group is taken as the reference category.

(4)

(iv) Write down the normal equations in part (ii) in terms of the model in part (iii) and hence solve them to find the parameter estimates for α, β and γ.

(7)