EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2017

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) Explain when a Latin Square design can be useful, and how it helps to satisfy the principles of experimental design.

(4)

(b) In an experiment at a horticultural research centre, material from 5 different suppliers (*A*, *B*, *C*, *D* and *E*) is used to study the growth of strawberry plants in a field which slopes down North to South and is also subject to a climatic trend (such as wind or rain) from East to West.

The research team wish to discover if the fertilisers supplied lead to significantly different quantities of strawberries. The following diagram shows the design used in the experiment.

		Column				
		1	2	3	4	5
Row	1	E	B	A	D	C
	2	C	A	D	E	B
	3	B	E	C	A	D
	4	A	D	B	C	E
	5	D	C	E	B	A

The response variable was a measure of the yield of strawberries on each plot, in suitable units. The row totals of the responses are 318, 330, 332, 339, 351, the column totals are 321, 331, 336, 337, 345, while the supplier totals are 331, 332, 393, 310 and 304 (all in the obvious order). The grand total is 1670.

An outline of the analysis of variance is as follows.

Source	d.f.	sum of squares	mean square	F ratio
Rows		118.0		
Columns		62.4		
Suppliers		994.0		
Residual			1.97	
Total	24	1198.0		

(i) Explain how these items have been calculated and complete the analysis of variance table.

(11)

(ii) Write a brief report on the results.

(5)

2. (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, the set (ε_i) being mutually independent with zero mean and constant variance.

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

(12)

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

(3)

(iii) It is suggested that the model given in part (i) ought to have a constant term (μ) added to it. Given that a suitable computer program for multiple regression is available, suggest two ways of comparing the model in part (i) with this modified model.

(5)

3. A production line makes small plastic components before they are transferred in batches of 500 to the inspection department. Twenty successive batches contained the following numbers of defective items:

18, 24, 27, 17, 36, 34, 15, 24, 21, 18, 30, 33, 19, 21, 20, 26, 32, 31, 21, 24.

(i) If no more than 5% of the items in a batch are defective, the whole batch is deemed acceptable. Draw a control chart which shows the data and the warning and action limits.

(11)

(ii) Use your chart to write a brief report on the component's production line.

(4)

(iii) Explain the use of cusum charts in this type of study, and what further information they can provide.

(5)

4. (a) A biologist has been studying how two different treatments affect the growth of a plant species. The treatments consist of different mixtures of fertiliser x, and measurements of growth y are taken after a fixed period of time.

Draw sketch graphs that illustrate each of the following 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 as 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 a, k are constants to be estimated, the relation $y = a\exp(kx)$ can be transformed to give a linear model. If the intercept in the linearised model has the point estimate 3.51, with 95% confidence interval (2.65 to 4.37), calculate the point estimate for a, and the corresponding confidence interval. Explain why the point estimate is not in the centre of the interval.

(6)

(ii) The following three models can be used to explain the dependence of a random variable y on x, or on x_1 and x_2, where y, x, x_1 and x_2 are strictly positive, and a, b, c, k are constants to be estimated. Give a linearised version of each of these models, and explain briefly how the parameters can be estimated.

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

$$y = \frac{x}{ax+b}$$

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

(6)