EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2011

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 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. (a) To compare two animal diets A and B, eight pairs of twin animals were used. One twin in each pair was chosen at random and given the diet A while the other twin received diet B. The gains in weights (kg) of the animals over the period of the experiment were as follows.

Pair	1	2	3	4	5	6	7	8
Diet A	25.6	20.7	14.0	21.5	21.6	25.6	26.6	22.2
Diet B	24.1	17.7	14.3	19.7	22.2	23.5	25.4	21.3

(i) Is there evidence of a difference between the effects of diets A and B on the mean weight-gains of the animals? State any assumptions that you make.

(9)

(ii) Explain briefly why this experimental design would be superior to one in which the animals given diet A were chosen independently of those given diet B.

(2)

(b) (i) Describe the circumstances in which a Latin square design is useful, and illustrate your answer with a practical example.

(3)

(ii) Write down the linear model to be used as the basis for an analysis of data from this design and state the assumptions required in it.

(3)

(iii) Give an outline of the analysis of variance table for a Latin square, listing the items in the analysis and their degrees of freedom.

(3)

2. In a randomised blocks experiment to determine people's reaction times to a stimulus (a flashing light) under different environmental conditions A – E, five students from the same age-group, all of whom had used the equipment before, recorded times (milliseconds) as shown in the following table.

		Student					
		I	II	III	IV	V	Total
Condition	A	213	127	155	246	200	941
	B	178	143	147	210	192	870
	C	254	151	174	266	222	1067
	D	103	108	122	144	161	638
	E	177	199	212	168	182	938
	Total	925	728	810	1034	957	4454

The sum of the squares of all 25 observations is 839 414.

(i) Write down the linear model you would use as a basis for analysing these data, giving the meaning of each item in it and specifying any assumptions that have to be made.

(3)

(ii) Carry out an analysis of variance for these data.

(11)

(iii) Obtain 95% confidence intervals for the difference between the means of conditions B and C, and for the difference between the means of conditions B and D.

(4)

(iv) What conclusions can you draw from these results regarding differences among all five conditions?

(2)

3. Small plastic components are made at one factory and then transported to a second factory where 100% inspection is carried out. The non-defective components are then used in a manufactured product. The transport between the two factories can handle batches of 500 at a time. One batch is transported each day.

On successive working days, 20 such batches gave the following numbers of defective items at the inspection stage.

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

(i) If there are less than 5% of defectives on average, the manufacturer who is going to use the components will not complain. Draw a control chart which shows the data and the warning and action limits.

(11)

(ii) Use your chart to report on the performance of the components' production line, and whether its operation should be checked frequently.

(4)

(iii) Suppose that a process is in control with exactly 5% defectives and a control chart with warning and action limits as in part (i) is in use. Find the probability of obtaining:

(a) two successive observations that are above the upper warning limit;

(b) two successive observations that are below the lower warning limit;

(c) three successive observations of which just the first and third are above the upper warning limit;

(d) three successive observations of which just the first and third are below the lower warning limit.

(3)

(iv) Processes are often controlled automatically by computer programs which inspect control charts as each new observation becomes available. One test used by such computer programs is to stop the process when any of the events in part (iii) occurs. Comment on the use of this test in comparison with standard use of Shewhart charts.

(2)

4. (a) In a set of data collected from 92 students (of similar ages), their heights and weights were measured. Heights were between 62 and 76 inches and weights between 95 and 215 pounds. A simple linear regression gave

$$\text{weight} = -205 + 5.09\ \text{height}, \qquad R^2 = 61.6\%.$$

Since the relationship between weight and height may be different for males and for females, a dummy (indicator) variable "sexM0F1" was added, coded 0 for males and 1 for females. A linear regression including this variable gave

$$\text{weight} = -103 + 3.69\ \text{height} - 14.7\ \text{sexM0F1}, \qquad R^2 = 66.1\%.$$

Explain the meaning of each of the coefficients in these two equations. Comment on how reasonable these equations are for explaining the relation between weight and height.

(7)

(b) In another set of data in which 5 variables (y, x_1, x_2, x_3, x_4) were recorded on 15 items, the regression of y on all four x variables gave $R^2 = 89.54\%$. The total (corrected) sum of squares was 1847.60. Values of R^2 for regressions on all subsets of the variables were as follows.

Variables included	R^2
$x_1,\ x_2,\ x_3$	89.44%
$x_1,\ x_2,\ x_4$	75.08%
$x_1,\ x_3,\ x_4$	89.53%
$x_2,\ x_3,\ x_4$	82.04%
$x_1,\ x_2$	74.67%
$x_1,\ x_3$	89.34%
$x_1,\ x_4$	75.02%
$x_2,\ x_3$	81.87%
$x_2,\ x_4$	13.56%
$x_3,\ x_4$	81.29%
x_1	74.33%
x_2	10.51%
x_3	81.28%
x_4	0.27%

Use the backwards elimination method to find a "best" subset of the x variables to explain y satisfactorily. Explain your working fully.

(13)