EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2014

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

There are 4 questions altogether in the paper.

1. A randomised blocks experiment is being run to investigate the effect of four different methods (A, B, C, D) of performing a task on the time taken to do it, using eight people of varying skill levels at the task. The times taken, in minutes, are given below.

	Person								
Method	1	2	3	4	5	6	7	8	Sum
A	32	30	32	27	27	32	31	36	247
B	29	38	33	31	36	41	36	35	279
C	30	40	32	30	37	35	32	34	270
D	31	36	30	31	33	33	37	32	263
Sum	122	144	127	119	133	141	136	137	1059

$\Sigma y^2 = 35\,407$

(i) Write down an appropriate linear model for these data, clearly stating any assumptions made.

(3)

(ii) Construct the analysis of variance table for these data. State appropriate hypotheses, and test at the 5% level whether there is any evidence of a difference in mean time taken between the methods. Clearly state your conclusions.

(9)

(iii) Construct appropriate 95% confidence intervals for the mean time taken for each of the methods. Explain what these intervals show and what they add to your answer to part (ii).

(4)

(iv) State any potential problems with this experiment, and comment on how it might be improved.

(4)

2. (a) A set of multiple regression models has been fitted to a data set with 16 data points and two input variables x_1 and x_2. Models with all appropriate combinations of linear, square and cross-product terms have been fitted, giving the residual sums of squares shown below.

Predictors	*Residual SS*	*Residual d.f.*
$x_1, x_2, x_1^2, x_2^2, x_1x_2$	47.8	10
x_1, x_2, x_1^2, x_2^2	63.1	11
x_1, x_2, x_1^2, x_1x_2	55.2	11
x_1, x_2, x_2^2, x_1x_2	61.4	11
x_1, x_2, x_1^2	64.3	12
x_1, x_2, x_2^2	71.8	12
x_1, x_2, x_1x_2	66.6	12
x_1, x_1^2	81.2	13
x_2, x_2^2	113.9	13
x_1, x_2	89.4	13
x_1	112.4	14
x_2	135.6	14
constant only	168.7	15

Use backwards elimination, with tests at the 5% level, to select the best regression model for these results, fully explaining and justifying your choice.

(10)

(b) Consider a case with two continuous explanatory variables x_1 and x_2. In a designed experiment to investigate their effects on a response variable y, both variables appear at three fixed levels, designated low, medium and high in each case. Each combination of these factor levels is observed twice.

One researcher claims that the results should be analysed using two-way analysis of variance, with interactions, while a second researcher claims that they should be analysed using multiple linear regression on the actual values of x_1 and x_2, with square and cross-product terms.

Explain the concept of indicator variables and show how the two-way analysis of variance model can be written as a regression model using indicator variables.

Hence compare and contrast these two methods of analysis, clearly stating the models used and the assumptions required in each case.

In particular consider the advantages and disadvantages of the two approaches in terms of flexibility, degrees of freedom and usefulness for prediction.

(10)

3. A trial is to be conducted to investigate the effects of different drugs and different exercise regimes on patients with a degenerative condition.

The trial will involve five different drug treatments, namely four new drugs and a placebo which will form the control group. It will involve four different exercise regimes, namely three different mandated forms of exercise and a regime with no mandated exercises to form the control group. All of the patients will be of similar age, both male and female.

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 do not need to go into the details of the analysis.

(i) Factorial designs.

(5)

(ii) Interactions.

(3)

(iii) Blinding and double blinding.

(4)

(iv) Randomisation.

(4)

(v) Blocking.

(4)

4. A company makes machine components and knows from past experience that the most important and difficult aspect of manufacture is to ensure that the components are of the right length to fit properly into the machines.

A component will fit properly if it is between 788 mm and 792 mm in length. If it is not, it will not fit and hence is deemed faulty.

(i) The company's current quality control method is regularly to take a sample of 100 components and calculate the proportion of faulty ones. Describe the basic principles of a Shewhart control chart and how in this case such a chart can be used to provide warning limits and action limits when it is desired that the proportion of faulty components be no higher than 10%.

(7)

(ii) The company is considering a new quality control method which instead takes samples of four components and calculates their mean length. Show how a control chart for the mean can be used to provide warning limits and action limits when it is desired that the component lengths are Normally distributed with mean 790 mm and standard deviation at most 1.21 mm.

(6)

(iii) Calculate the probability of a sample lying outside an action limit in part (i), and the probability of a sample lying outside an action limit in part (ii), in the case when individual component lengths are Normally distributed with mean 791 mm and standard deviation 1 mm.

(4)

(iv) Briefly compare and contrast the relative advantages and disadvantages of the two methods in parts (i) and (ii), in general and in this particular case.

(3)

BLANK PAGE

BLANK PAGE

BLANK PAGE