EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2008

Paper III : Statistical Applications and Practice

Time Allowed: Three Hours

*Candidates should answer **FIVE** questions.*

All questions carry equal 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 $\begin{pmatrix} n \\ r \end{pmatrix}$ *is the same as* $^{n}C_{r}$.

This examination paper consists of 10 printed pages **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. Sam is keen on doing "Sudoku" puzzles, but has so far only attempted ones labelled as "Easy" or "Moderate". Moderate puzzles are considered to be harder than Easy puzzles but Sam is not sure that on average there is a genuine difference in difficulty. Equating difficulty with the length of time that a puzzle takes to complete, he decides to carry out an experiment to assess whether Moderate Sudokus take longer to complete on average than Easy Sudokus.

   Sam selects random samples of 15 Easy Sudokus and 15 Moderate Sudokus, from a book of a large number of Sudokus, and solves one each day for 30 days, with the puzzle for each day chosen at random. He records the number of minutes he takes to complete each puzzle successfully.

   The times (in minutes) taken to complete the puzzles are as follows, sorted into ascending order of magnitude purely for computational convenience.

Easy	x	9, 10, 10, 12, 12, 13, 14, 15, 15, 18, 20, 20, 22, 23, 24
Moderate	y	13, 14, 16, 18, 18, 18, 18, 18, 19, 19, 21, 23, 24, 24, 24

   You are given the following summary statistics, where $\{x_i\}$ are the times to do the Easy puzzles and $\{y_i\}$ are the times to do the Moderate puzzles.

   $$\sum_{i=1}^{15} x_i = 237 \qquad \sum_{i=1}^{15} x_i^2 = 4097 \qquad \sum_{i=1}^{15} y_i = 287 \qquad \sum_{i=1}^{15} y_i^2 = 5661$$

   (i) Carry out a two-sample t test, stating carefully your null and alternative hypotheses.

   (10)

   (ii) State why you might consider a non-parametric test to be appropriate here. Re-analyse the data using a Wilcoxon test, once again stating carefully your null and alternative hypotheses.

   (8)

   (iii) Comment on your conclusions in parts (i) and (ii), and also highlight any reservations you might have about the design of this experiment.

   (2)

2. (i) What are the essential features of a factorial experiment? State one advantage and one disadvantage of such an experiment, compared with an experiment in which factors are examined one at a time.

(5)

(ii) The effectiveness of a particular training programme is considered to be dependent on the training medium and the training location. A 2^2 factorial experiment was run in which the factors were training medium [virtual learning environment (VLE) or face-to-face (F2F)] and training location [local or central]. One hundred subjects were randomly divided into four groups of 25 and each group was assigned to one of the four treatment combinations. Each individual was given a test at the end of the training session and the mean score for each group is shown in the table below.

		Training	
		VLE	F2F
Location	*Local*	120	76
	Central	84	80

(a) Estimate the training and location main effects and the training $\times$ location interaction.

(6)

(b) State the number of degrees of freedom of the estimator of the residual variance. The estimated residual variance is 160. Test the null hypothesis that the interaction effect is zero against a two-sided alternative. Explain how the presence of an interaction has implications for the estimation of main effects.

(7)

(iii) Comment on a possible deficiency of this experiment.

(2)

3. The Weibull distribution has two parameters $a > 0$ and $b > 0$ and has cumulative distribution function (cdf)

$$F(x) = 1 - \exp\left\{-\left(\frac{x}{a}\right)^b\right\}, \quad x > 0.$$

(i) Show that the probability density function is

$$f(x) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} \exp\left\{-\left(\frac{x}{a}\right)^b\right\}, \quad x > 0.$$

(4)

(ii) Taking the value of b to be fixed, show that the maximum likelihood estimate for a, based on a random sample of observations $x_1, x_2, \ldots, x_n$ from a Weibull distribution, is given by

$$\hat{a} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^b\right)^{1/b}.$$

(8)

(iii) From past experience it is known that the lives of ball bearings, measured in millions of revolutions, follow a Weibull distribution with $a = 75$ and $b = 3$. However, after a change in production process, it is thought that b should remain unaltered but that the value of a might have changed.

A random sample of 12 ball bearings was tested to failure and the lifetimes at failure, in millions of revolutions, are given below.

17.88   33.00   42.12   48.48   51.96   55.56   68.64   68.88   93.12
105.12   127.92   173.40

Calculate the maximum likelihood estimate of a, given that $\sum_{i=1}^{12} x_i^3 = 10\,468\,316$.

(2)

(iv) The plot below shows the empirical cdf (Femp) and the fitted cdf (Ffit), using $b = 3$ and the maximum likelihood estimate for a. Comment on the fit of the proposed model. Suggest two possible courses of action to obtain a model with improved fit.

(6)

The plot is on the next page

Turn over

Scatterplot of Ffit, Femp vs x

Turn over
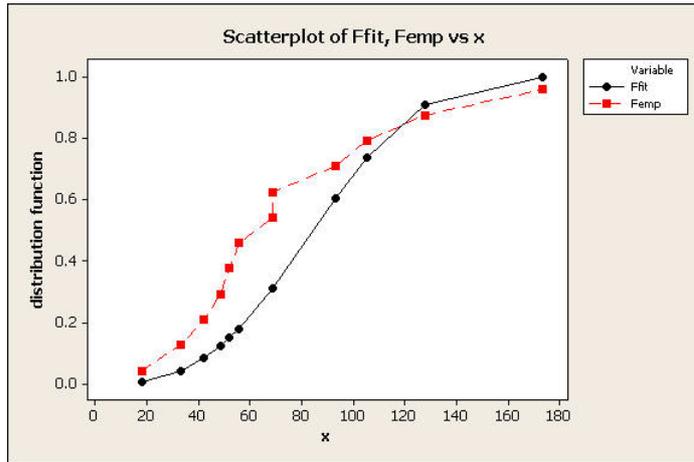
4.   A School of Business Studies wishes to investigate whether the proportion of female students on postgraduate courses varies across different areas of provision. Of particular interest is a comparison between students who attend Master of Science courses (MSc) and those who attend Master of Business Administration courses (MBA). A current cohort of students is taken as a sample cohort, and it is considered that the cohort is typical of other recent cohorts at this School and of cohorts at similar Business Schools.

Numbers of students in the different categories are as follows.

	MSc	MBA
Male	716	169
Female	466	51

(i)   Carry out a χ^2 test to determine whether or not there is any association between the type of course studied and the sex of students.

(5)

A further question of interest is whether the subject matter of the postgraduate courses is relevant to the level of female participation. The MSc courses are either finance or management oriented. The MBA is predominantly a management course. The MSc students are generally younger than the MBA students, who have work experience and are often sponsored by their employers. Subdividing the MScs into finance and management gives the following table.

|	MSc (finance)	MSc (management)	MBA (management)
Male	611	105	169
Female	363	103	51

(ii)   Analyse these data in a manner that allows you to comment on the relationship between the type of Master's course studied and the proportion of female students, and also to comment on any influence of the type of subject matter. Write a short report (5 or 6 sentences) to summarise your findings.

(15)

Turn over

5. Answer the following in the context of sample surveys.

(i) Define *simple random sampling*. State two properties that must hold as each member of the sample is selected.

(4)

(ii) Describe *systematic random sampling*. Contrast this process with simple random sampling, and highlight any advantages of the process and any potential problems with the process.

(5)

(iii) Describe the circumstances in which it might be desirable to use stratified sampling.

(3)

(iv) Describe the difference between sampling error and non-sampling error, illustrated by two examples of each type of error.

(6)

(v) The aims of survey planning are to minimise both costs and errors. Discuss briefly why these requirements conflict.

(2)

7

6. The following table shows the quarterly imports, *Y*, of fruit (in tonnes) handled by an expanding marketing company for the years 2004 to 2007.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
2004	19	11	5	13
2005	27	18	7	22
2006	40	20	11	26
2007	55	37	16	45

(i) A time series plot of the data is given below. Describe the main features of the time series. Explain why an additive model will not adequately account for the data. (3)
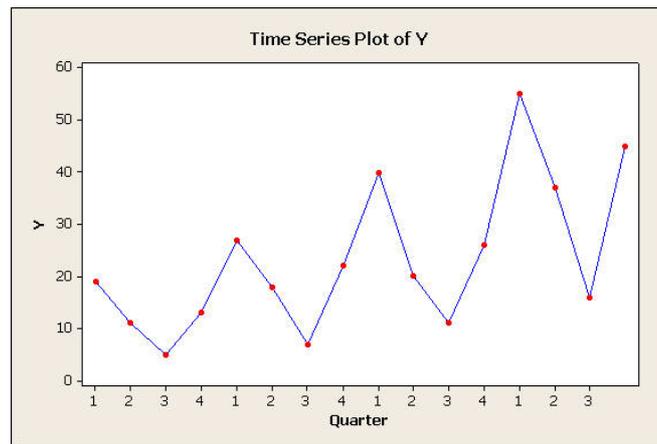
(ii) A multiplicative model of the form $Y_t = T_t S_t I_t$ is chosen, where *T* is the trend, *S* is the seasonal component and *I* the irregular (residual) component. Show how this model can be represented in additive model form. (1)

(iii) Describe the moving average formula used in the table below and explain why it is a suitable choice here. Use the differences "$\log Y - MA$" to estimate $\log S$ for each of the four quarters. Hence obtain the values of $\log I$ in the quarters from 2004:3 to 2007:2. Draw a time series plot of $\log I$ and comment on how well the model fits the data.

```
Year    Qtr      logY        MA      logY-MA
2004     1     2.94444
         2     2.39790
         3     1.60944    2.42311    -0.81367
         4     2.56495    2.52859     0.03636
2005     1     3.29584    2.63221     0.66363
         2     2.89037    2.74003     0.15034
         3     1.94591    2.85492    -0.90901
         4     3.09104    2.91722     0.17382
2006     1     3.68888    2.98689     0.70199
         2     2.99573    3.06427    -0.06854
         3     2.39790    3.12496    -0.72706
         4     3.25810    3.24166     0.01643
2007     1     4.00733    3.36540     0.64194
         2     3.61092    3.48080     0.13011
         3     2.77259
         4     3.80666
```

(13)

(iv) Discuss the relative merits of estimation of the trend in a time series using (a) moving averages, and (b) linear regression. (3)

8

7. (a) A quality control sampling scheme for large batches of a mass-produced product operates as follows.

A simple random sample of 25 units is drawn from a batch. If the number of defective units is zero, the batch is accepted. If the number of defective units is 2 or more, the batch is rejected. If the number of defective units is 1, then a second random sample of 25 is taken from the batch, independently of the first sample. If the number of defective units in the second sample is zero then the batch is accepted, otherwise the batch is rejected.

   (i)   If the batch contains a proportion p of defectives, calculate the probability of accepting a batch for $p = 0.005, 0.01, 0.02$.

   (8)

   (ii) For each value of p in (i), calculate the mean sample size for this scheme.

   (4)

   (iii) It is suggested that the second stage of the scheme should, instead, consist of inspection of the entire batch, followed by rejection of any defective units and acceptance of the rest. Comment on this suggestion.

   (2)

   (b) A survey is to be conducted amongst students at a particular university to estimate the mean number of hours spent in part-time work per week.

   (i)   Previous studies of employment patterns of students at universities have established 9.3 hours as the standard deviation of hours worked. Calculate how many students, approximately, should be sampled in order to estimate the mean to within 1 hour, at a 95% level of confidence. You may assume in your calculation that the population standard deviation is equal to the above figure.

   (4)

   (ii) What additional questions would you ask about the data collected, with a view to clarifying the information about the employment pattern of students?

   (2)

9

8. In a completely randomised experiment designed to compare three different treatments against a control, treatment T_j is assigned to n_j experimental units, $j = 0, 1, 2, 3$, where $j = 0$ is the control. It is assumed that observations come from $N(\mu_j, \sigma^2)$ distributions.

(i) What do you understand by the term *contrast*?

What is the contrast $\theta = \mu_0 - \frac{1}{3}(\mu_1 + \mu_2 + \mu_3)$ designed to measure?

(3)

(ii) An experiment is carried out and gives the following results.

		Sample mean	Sample variance
T_0	18.89, 17.84, 14.93, 16.31, 13.17, 14.80, 13.07, 17.48, 14.81, 18.14, 14.92, 14.68, 16.49	15.81	3.49
T_1	18.83, 15.59, 20.21, 16.21	17.71	4.75
T_2	22.60, 17.42, 17.95, 17.38, 20.60	19.19	5.39
T_3	19.72, 21.62, 20.17, 17.27, 14.52	18.66	7.81

Copy and complete the following analysis of variance table and state your conclusions.

Source	DF	SS	MS	F
Treatment	****	****	****	****
Residual	****	****	****	
Total	****	165.49		

(8)

(iii) Write down an estimate of the contrast θ in part (i), and calculate an estimate of the variance of the estimator of θ. On the basis of these values and the analysis in part (ii), do you consider that there are significant treatment effects with respect to the control? Briefly justify your answer.

(4)

(iv) Suppose now that there are sufficient resources to use N experimental units in total, and that the same number of units, n, is allocated to each of the three treatments. The control is thus allocated to $N - 3n$ units. Find an expression for n that minimises the variance of the estimated θ. Calculate how this would have been applied to the above experiment if there had been a total of 30 experimental units available.

(5)