EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2003

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 silent, cordless, non-programmable electronic calculators.

Where a calculator is used the ***method*** *of calculation should be stated in full.*

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 9 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. Three types of watch dial were tested on 21 subjects under simulated conditions. One dial was assigned at random to each subject and the number of errors the subject made in reading this dial during a standardised series of tests was recorded. The results are shown in Table 1 below.

Table 1

Dial type		
1	*2*	*3*
42	62	56
30	53	36
21	61	43
47	47	58
34	45	46
22	59	24
42		31
38		

Incomplete results of a one-way analysis of variance of the data are shown in Table 2.

Table 2

One-way ANOVA: type 1, type 2, type 3

```
Analysis of Variance
Source     DF        SS
Dial type   2      1377
Error      18      1858
Total      20      3234
```

(i) Complete the analysis and interpret your results, stating any assumptions you have made in reaching a conclusion.

(6)

(ii) Estimate the difference between the mean numbers of errors that would be made by subjects reading dials of type 1 and of type 2, and find a 95% confidence interval for this difference. Explain what is meant by describing your interval as a "95% confidence interval". You may take it that any assumptions needed for your analysis are satisfied.

(10)

(iii) How could you investigate the assumptions needed for the one-way analysis of variance of the data in Table 1? If you were unwilling to accept these assumptions, explain briefly how you might proceed. (Do not actually do so.)

(4)

2. An experiment, described in the *Journal of Materials Science 1986*, was conducted to investigate the effect of antimony on the strength of tin-lead solder joints. Four different amounts of antimony were considered (0%, 3%, 5% and 10% weight) and four different cooling methods. Each of the four different amounts of antimony was used with each of the cooling methods, giving a total of sixteen different treatments. Three joints were assigned at random to each of these sixteen treatments and their shear strengths were measured in MPa.

Table 1 shows the mean shear strength at each treatment. Table 2 shows some results of an analysis of variance of the data.

Table 1. Mean shear strengths

```
Rows: amount of antimony (% weight)
Columns: cooling method

          AB        FC        OQ        WQ       All

 0    20.333    19.833    22.067    18.467    20.175
 3    22.233    19.933    20.433    19.033    20.408
 5    21.400    18.833    21.467    20.767    20.617
10    16.733    17.200    17.833    16.300    17.017

All   20.175    18.950    20.450    18.642    19.554

Key: AB = air-blown
     FC = furnace-cooled
     OQ = oil-quenched
     WQ = water-quenched
```

Table 2

Two-way ANOVA: strength versus cooling, antimony

```
Analysis of Variance for strength

Source         DF        SS         MS        F        P
Cooling         3     28.63       9.54     5.53    0.004
Antimony        3    104.19      34.73    20.12    0.000
Interaction     9     25.13       2.79     1.62    0.152
Error          32     55.25       1.73
Total          47    213.20
```

(i) Explain what is meant by *interaction*, both generally and in the context of this experiment.

(6)

(ii) By referring to the results of the analysis of variance, and by drawing a suitable plot, comment on whether there appears to be interaction between the amount of antimony and the cooling method.

(7)

(iii) Explain carefully how to interpret the F and P values for cooling and for antimony in Table 2, stating any assumptions needed for your answer.

(7)

3. A random sample of n insurance policies of a particular type, all of which have been in force for 10 years, has been taken. The numbers of claims made are $x_1, x_2, \ldots, x_n$. A Poisson distribution with parameter λ is to be used to model the number of claims.

(i) Derive the maximum likelihood estimator of λ.

(5)

(ii) An insurance company has taken a random sample of 75 policies of the same type. All of these have been in force for ten years. The numbers of claims made are summarised below.

Number of claims	*Number of policies*
0	18
1	25
2	13
3	10
4	6
5	3
	Total 75

(a) Use an appropriate statistical test to assess whether a Poisson distribution is an appropriate model for this set of data.

(10)

(b) Obtain an approximate 95% confidence interval for the mean of the underlying distribution.

(5)

4. The table below shows UK households' final consumption expenditure on alcohol and tobacco, in £millions at 1995 prices, for the years 1990 to 2000 inclusive. (Source: Economic Trends Annual Supplement.) The table also shows the forecasts obtained from exponential smoothing, using 0.8 as the smoothing constant and taking the 1989 value as the forecast for 1990, and the errors in the forecasts.

Year	*Expenditure*	*Forecast*	*Error*
1990	20730	20735	–5
1991	20148	20731	–583
1992	19539	20265	–726
1993	19255	19684	–429
1994	19268	19341	–73
1995	18776	19283	–507
1996	19299	18877	422
1997	19459	19215	244
1998	19193	19410	–217
1999	19863	19236	627
2000	19959	19738	221

(i) Explain, in a non-technical way, how to calculate the forecasts shown here. Illustrate your answer by showing the calculation of the forecast for 2000 in detail.

(6)

(ii) When, in general, is it appropriate to use a high value for the smoothing constant?

(2)

(iii) Describe one method of measuring the accuracy of the fitted model. How would this measure usually be used in practice?

(5)

(iv) Use the given results to forecast expenditure in 2001.

(2)

(v) Plot the errors in the forecasts and comment.

(5)

5. In World War II, a particular shipyard built numerous Liberty Ships. Orders for ships were given serial numbers, so the first such order was given the order number (NBR) 1, the second such order was given NBR 2, and so on. The number of thousand man-hours per ship built to meet each order (HRS) was noted. Table 1 below shows a number of values of NBR together with the corresponding values of HRS. Table 2 gives some regression results.

It is conjectured that man-hours per ship depends on the order number.

(i) Draw a graph showing the data in Table 1. Is the relation between man-hours per ship and order number linear? Justify your answer. (8)

(ii) On the basis of the regression results, which of the three linear regressions reported do you consider is best? Justify your answer. (2)

(iii) Interpret the estimated constant and regression coefficient for the regression you considered to be best in part (ii). (8)

(iv) What other transformation of the data might you try to improve further on the regression you considered to be best, and why? (2)

Table 1

NBR	5	10	25	30	75	100	125	150	175	200	225	250	285
HRS	1094	894	647	659	529	424	376	395	395	388	376	353	280

Table 2

Regression Analysis: HRS versus NBR

```
The regression equation is
HRS = 790 - 2.09 NBR

Predictor        Coef     SE Coef         T        P
Constant       789.54       65.77     12.00    0.000
NBR           -2.0871      0.4190     -4.98    0.000

S = 138.8       R-Sq = 69.3%     R-Sq(adj) = 66.5%
```

Regression Analysis: HRS versus 1/NBR

```
The regression equation is
HRS = 391 + 3975 1/NBR

Predictor        Coef     SE Coef         T        P
Constant       391.14       27.34     14.31    0.000
1/NBR          3975.2       427.4      9.30    0.000

S = 84.09       R-Sq = 88.7%     R-Sq(adj) = 87.7%
```

Regression Analysis: HRS versus logNBR

```
The regression equation is
HRS = 1306 - 181 logNBR

Predictor        Coef     SE Coef         T        P
Constant      1306.24       48.14     27.13    0.000
logNBR        -180.51       10.67    -16.92    0.000

S = 48.17       R-Sq = 96.3%     R-Sq(adj) = 96.0%
```

6. Consider the following three situations (I), (II) and (III) concerned with surveys of subscribers to two journals A and B. Each journal had a very large number of subscribers.

(I) *In a random sample of 150 subscribers to journal A, 84 agreed with the statement "This journal is very useful to me in my work". In a random sample of 200 subscribers to journal B, 126 agreed with the same statement.*

(II) *In the same random sample of 150 subscribers to journal A, 45 agreed with the statement "This journal is quite useful to me in my work" and 21 agreed with the statement "This journal is not at all useful to me in my work".*

(III) *In the same random sample of 200 subscribers to journal B, 43 agreed with the statement "The current issue of this journal is not at all useful to me in my work" and 56 agreed with the statement "The previous issue of this journal is not at all useful to me in my work".*

(i) Obtain an approximate 95% confidence interval for the difference between the proportion of subscribers to journal A agreeing with the statement "This journal is very useful to me in my work" and the proportion of subscribers to journal B agreeing with this statement (situation I). Comment.

(7)

(ii) Explain carefully why the method used in part (i) is not appropriate for obtaining a confidence interval for the difference between the proportion of subscribers to journal A agreeing with the statement "This journal is quite useful to me in my work" and the proportion of subscribers to journal A agreeing with the statement "This journal is not at all useful to me in my work" (situation II).

(4)

(iii) Explain carefully why the method used in part (i) is not appropriate for obtaining a confidence interval for the difference between the proportion of subscribers to journal B agreeing with the statement "The current issue of this journal is not at all useful to me in my work" and the proportion of subscribers to journal B agreeing with the statement "The previous issue of this journal is not at all useful to me in my work" (situation III).

(4)

(iv) On the basis of the sample result given in situation I above, how large a sample of subscribers to journal B would be needed in order that the investigators would have 95% confidence that the estimate of the proportion of subscribers to journal B agreeing with the statement "This journal is very useful to me in my work" is no more than 0.05 different from the true proportion?

(5)

Turn over

7. (i) (a) Explain why non-response is a problem in social surveys.

(4)

(b) Suggest how non-response might be reduced in a mail survey by appropriate follow-up procedures.

(4)

(ii) An interview survey of heads of household is to be undertaken and it has been estimated that 600 completed interviews are needed.

(a) Comment on the following two strategies (A) and (B) for achieving 600 completed interviews.

(A) *Give the team of interviewers contact details of a large sample of heads of household and give instructions that interviewing should stop once 600 interviews have been completed.*

(B) *Give the team of interviewers contact details of a sample of 600 heads of household and give instructions that only one attempt is to be made to interview each of these heads. If no interview is obtained from m of these 600 heads of household then give the team contact details of a further sample of m heads of household and give instructions that several attempts are to be made to interview the members of this further sample.*

(7)

(b) A simple random sample of n heads of household is to be selected. It is thought that the probability of any particular head of household responding is 0.75. What is the smallest value of n such that there is a probability of at least 0.99 of obtaining 600 or more responding households?

(5)

8. The table shows details of employees in the UK by sex and industry at June for four selected years. Performing such calculations on the data as you think appropriate, write a short report outlining the main features shown in this table, supporting your answer by diagrams. Give particular attention to differences in trends between males and females.

Note: You are advised to show only a selection of the data in diagrams.

Percentages

	Males				*Females*			
Industry	*1978*	*1981*	*1991*	*1997*	*1978*	*1981*	*1991*	*1997*
A. Distribution, hotels, catering and repairs	15	16	19	20	24	25	25	26
B. Manufacturing	35	33	26	26	22	18	12	10
C. Financial and business services	9	10	15	16	11	12	16	19
D. Transport and communication	9	9	9	9	3	3	3	3
E. Construction	8	8	8	7	1	1	1	1
F. Agriculture	2	2	2	2	1	1	1	1
G. Energy and water supply	5	5	3	1	1	1	1	–
H. Other services	16	17	19	19	38	39	41	40
All employees (=100%) (millions)	13.4	12.6	11.5	11.5	9.4	9.3	10.7	11.3

Source: Short-term Turnover and Employment Survey, Office for National Statistics. Reproduced in Social Trends 28 Pocketbook