EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2003

Paper II : Statistical Methods

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. (i) A sports equipment company has commissioned an advertising agency to develop an advertising campaign for one of its new products. They can choose between two particular television commercials, *A* and *B*. To aid them in their decision, an experiment is performed in which 200 volunteers are randomly assigned to view one of the two commercials, 100 being assigned to each. After seeing the commercial, each volunteer is asked to state whether they would consider buying the product, with the following results.

		Commercial	
		A	*B*
Purchase product	*No*	70	80
	Yes	30	20

Apply a chi-squared test to these data and comment on your results. What recommendations, if any, would you make to the sports manufacturer concerning the choice of commercial for the proposed advertising campaign?

(7)

(ii) A random sample of sportswear manufacturers was surveyed to determine whether they advertised on television and/or the internet. The results are given in the following table.

		Internet	
		No	*Yes*
Television	*No*	3	5
	Yes	15	17

Apply McNemar's test to the above data and comment on your results.

(7)

(iii) Distinguish carefully between chi-squared tests and McNemar's tests, as used to analyse data such as given in parts (i) and (ii) of this question, giving examples of when each would be preferred to the other.

(6)

2. The telephonist answers telephone calls arriving at the switchboard of a particular organisation. A random sample of 100 calls received at the switchboard on a particular day was monitored and the time taken for the telephonist to answer was recorded. The data obtained are summarised in the following table.

Time in seconds	**Number of Calls**
< 10	5
≥ 10 but < 20	16
≥ 20 but < 25	10
≥ 25 but < 30	20
≥ 30 but < 35	21
≥ 35 but < 40	14
≥ 40 but < 50	10
≥ 50 but < 70	4
Total	100

(i) Draw a histogram depicting the above data.

(7)

(ii) Estimate the mean and median of the data. What do the data and your statistics indicate about the distribution of the number of seconds it takes for the telephonist to answer a call?

(6)

(iii) Construct a 95% confidence interval for the mean number of seconds for a call to be answered, stating any assumptions that you make.

(7)

3. One of the tasks routinely undertaken by a particular laboratory is to establish the potassium content of blood serum. It has been established that, when apparatus is working properly, in repeated tests of the same blood serum the standard deviation should not exceed 0.05g (%).

The manager of the laboratory decides to perform a quality control study of the two sets of apparatus used by the laboratory to measure the potassium content of various compounds. A test sample is prepared in which the potassium content is known to be 10.5g (%) and each set of apparatus is used to make eight repeat analyses of the test sample. The results in g (%) are as follows.

Apparatus A	10.55	10.62	10.40	10.52	10.46	10.31	10.50	10.49
Apparatus B	10.30	10.25	10.35	10.30	10.28	10.35	10.24	10.43

For each set of apparatus is there significant evidence that

(i) the readings are more variable than the general laboratory standard?

(8)

(ii) the readings are biased?

(8)

Comment on how each set of apparatus should be altered to improve the accuracy of its measurements.

(4)

Turn over

4. (i) A psychologist claims that visual memory is more effective than aural memory. To test this claim, ten students are selected at random and examined for visual and aural memory using a standard memory test. For each student, the psychologist notes whether his or her aural (*A*) or visual (*V*) memory score is the greater. The results are as follows.

Student	1	2	3	4	5	6	7	8	9	10
Test	*A*	*V*	*V*	*A*	*V*	*V*	*V*	*A*	*V*	*V*

Carry out a suitable analysis of these data to investigate the psychologist's claim and comment on your results.

(7)

(ii) After completing this experiment, the psychologist decides to investigate whether aural memory scores can be improved by coaching. A second experiment is conducted in which a random sample of 12 students perform an aural memory test before and after several sessions of coaching in skills believed to aid aural memory. The results of the two tests are as follows.

Before	53	59	61	48	39	56	75	45	73	60	69	66
After	60	57	67	52	63	71	70	46	76	65	62	65

(a) It is required to test whether aural memory is improved, using a non-parametric test. Explain why it is not satisfactory to use a test similar to the one used in part (i). Carry out an appropriate non-parametric test.

(8)

(b) It is suggested that a parametric test would be more appropriate to analyse the data in (ii). Without performing the analysis, state which test you would use and any assumptions necessary for this analysis to be valid. Would these assumptions be reasonable in this case?

(5)

5. (i) Explain informally the *central limit theorem* and briefly explain its practical importance.

(6)

(ii) A pharmaceutical company needs to determine whether a new drug is effective in the treatment of elderly patients with insomnia. To investigate this, 288 patients over the age of 65 suffering from insomnia were randomised to receive the new drug or a placebo for a period of 6 weeks. At the end of this period, each patient was asked to complete a diary card for the next seven days, indicating the number of hours they had slept the preceding night. The total number of hours slept by each patient during this assessment period was then recorded. The following table gives summary statistics for the number of hours slept during the assessment period for the patients in each of the two groups.

	New Drug	*Placebo*
Number of patients	144	144
Mean number of hours slept per patient	50.6	35.4
Variance of number of hours slept per patient	10.3	14.7

Construct a 95% confidence interval for the difference in the mean number of hours slept per patient between the two groups and interpret your findings.

(7)

(iii) In a trial of anti-inflammatory drugs in the treatment of eczema, each member of a sample of 500 adults suffering from eczema was allocated at random to receive one of two treatments. After one month, the patients were asked to state whether their eczema improved. They replied as follows.

	Improved	*Not Improved*
Treatment A	205	45
Treatment B	180	70

Construct an approximate 95% confidence interval for the difference in the proportion of eczema sufferers in the population who would report an improvement if given treatment *A* rather than treatment *B*.

(7)

6. (i) State and explain a linear model that can be used as the basis for a one-way analysis of variance. Explain clearly what each term in the model represents and state any assumptions required for the analysis to be valid.

(6)

(ii) (a) A farmer is considering which of a range of fertilisers to use on his potato crop. To help him decide, he set up an experiment in which a field containing seed potatoes was divided into 20 equal plots, and one of four fertilisers *A*, *B*, *C* or *D* was randomly allocated to each plot, as in the diagram below, in which north is at the top.

D	A	D	A	B
C	C	A	B	B
A	D	C	D	C
C	B	D	B	A

Having grown the crop, the yield of potatoes, in kilograms, obtained from each plot was recorded as follows.

A	*B*	*C*	*D*
25	30	23	23
20	33	25	28
21	32	26	26
22	35	24	23
24	31	21	24

Carry out a suitable analysis of these data and write a report for the farmer, who is not trained in statistics, clearly stating your recommendations about which fertiliser he should use if he wants to maximise his potato yield.

(10)

(b) The amount of water in the soil can affect the yield of potato crops. If the farmer had suspected that the drainage in the field varies in an east-west direction, how might the design of the experiment be altered to take account of this?

(4)

7. The table below is derived from Table 6.1 of the report on the Family Expenditure Survey 2000 – 2001. It shows household expenditure in the United Kingdom during 1990 – 2001. All expenditure figures are shown at 2000 – 2001 prices. Using the information in this table, write an article for a serious newspaper on the distribution of and trends in household expenditure during this time. Your article should incorporate such diagrams and such statistics calculated from the table as you think appropriate.

(20)

Household expenditure 1990 to 2000-01, at 2000-01 prices.

Year	1990	1992	1994–95	1995–96	1996–97	1997–98	1998–99	1999–00	2000–01
Commodity or service	**Average weekly household expenditure (£)**								
Housing	60.40	58.60	54.70	55.60	54.60	55.50	59.80	58.70	63.90
Fuel and Power	15.10	16.10	15.30	14.60	14.80	13.50	12.20	11.70	11.90
Food and Drink	74.50	72.70	73.90	75.70	77.20	76.90	76.20	77.20	76.90
Tobacco	6.50	6.70	6.60	6.70	6.90	6.80	6.10	6.20	6.10
Clothing and Footwear	21.80	20.30	20.20	20.30	20.90	21.90	22.70	21.60	22.00
Household goods and services	43.90	43.70	44.50	44.50	47.70	48.20	50.80	51.10	54.60
Personal goods and services	12.90	12.60	12.70	13.40	13.20	13.70	13.90	14.30	14.70
Motoring and travel	54.40	53.00	50.40	51.20	55.60	60.30	62.80	63.50	64.60
Leisure goods and services	44.60	50.60	53.10	53.90	56.60	61.30	62.50	64.40	70.30
Miscellaneous	1.90	2.20	2.70	1.40	1.10	1.20	1.30	1.50	0.70
Total	335.80	336.40	334.00	337.30	348.50	359.20	368.40	370.20	385.70

Note: entries may not add to totals because of rounding.

Turn over

8. (i) Using examples to illustrate your answers, discuss the uses of the F distribution in statistical methods.

(8)

(ii) A manufacturer of soft drinks needs to purchase a new bottle-filling machine. The purchasing department has identified two machines, A and B, that are identical with respect to size, cost and convenience. The deciding factor will be the variability in the amount of liquid delivered by each machine, with the machine having the lower variability being preferred. Prior to purchase, the soft drinks manufacturer negotiates the chance to have each machine on one week's trial. During this week, the normal production run is shared between the two machines, each being set to produce on average one litre of soft drink per bottle. A random sample of bottles obtained from the production run of each machine is examined and the actual quantities of soft drink produced recorded in millilitres as follows.

Machine A

1002 1002 1001 995 1000 1007 1001 1004 998 993 1002 999
998 999 1000 1001

Machine B

1002 998 1001 1001 1000 999 1002 1000 1002 1000 1000 997
1001 1000 996 1001 1000 999 999 1003

Using a suitable statistical test, compare the variabilities in the amounts delivered by the two machines and hence write a short report to the manufacturer stating your recommendations concerning the choice of purchase.

(12)