EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2005

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 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 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. The following table gives a summary of the numbers of goals scored by the home soccer teams in matches in the English Premier Football League during the 1999–2000 season. It is required to test the assumption that the data follow a Poisson distribution.

Number of goals scored by the home team, r	0	1	2	3	4	≥ 5
Frequency, f	81	112	101	44	28	14

(i)   Explain why it might be reasonable to assume that the number of goals, r, scored by the home team would follow a Poisson distribution.

(3)

(ii) The total number of matches was $\Sigma f = 380$, and the total number of goals scored was $\Sigma fr = 634$. Also $\Sigma fr^2 = 1778$. Calculate the mean and variance of the data.

(3)

(iii) Calculate the expected frequencies, on the Poisson hypothesis, for $r = 0$ and $r = 1$. The expected frequencies in the remaining cells of the table are 99.72, 55.46, 23.13 and 10.51. Carry out a χ^2 goodness-of-fit test of the hypothesis that the data follow a Poisson distribution. Explain your conclusions carefully. What problem in carrying out the test would have occurred if the frequencies for values of $r \geq 5$ had not been combined?

(11)

(iv) What distribution would you have tried fitting to the data if the variance had been considerably larger than the mean? Briefly explain your reasoning.

(3)

2

2.   A hospital is interested in whether its stroke admissions are using more than the recommended average bed occupation time of 14 days in acute care for those discharged alive. The hospital audit department supplies you with bed occupation data from a random sample of 23 such discharges, as follows (in days).

44	20	7	19	14	6	5	3	8	6	6	5
1	4	2	9	17	1	3	7	82	4	19	

(i)   Find the median (M) and the quartiles (Q_1 and Q_3) of the data. Using the convention that any observation lying further than $1.5(Q_3 - Q_1)$ beyond the nearest quartile is an "outlier", draw a box and whisker plot of the bed occupation data. Hence comment on the shape of the distribution.

(8)

(ii)   The audit department wishes to have a test of the hypothesis that the *mean* duration of bed occupation is not greater than 14 days. Give a brief justification of the need for a larger sample of data in order to make a valid test of this hypothesis.

(3)

(iii)   The audit department now supplies you with a sample of 100 such admissions, for whom the duration of bed occupation can be summarised by

$$\Sigma x = 1488, \quad \Sigma x^2 = 44632.$$

Test the null hypothesis that the mean duration of bed occupation is 14 days against the alternative that it exceeds 14 days.

Discuss critically how reliable this result may be, and whether it is a useful measure of the hospital's performance.

(9)

3

Turn over

3.   In a small survey of perceived health risks in the UK, each member of a random sample of 50 people was asked the question "When buying food, do you check the pack for artificial additives?". The researchers wanted to discover whether females or males were more likely to check for artificial additives when buying food.

		Sex	
		Female	*Male*
Answer	*No*	18	17
	Yes	11	4

(i)   Test for a difference between the percentages of males and females responding "Yes" to the question about checking for artificial additives.

(9)

(ii)   Calculate an approximate 95% confidence interval for the difference in percentages of males and females responding "Yes" to the question about checking for artificial additives. How good do you believe the approximation to be? (State your reason.)

(11)

4

Turn over

4. The table below appeared in a report on the use of serologic screening of blood samples for toxoplasmosis (*Toxoplasma gondii*). The data are the results of two tests, the microscopic agglutination test (MAT) and the enzyme-linked immunosorbent assay (ELISA), on blood samples from 462 pigs.

		ELISA	
		Positive	*Negative*
MAT	*Positive*	67	25
	Negative	41	329

Source: Georgiadis M.P., Johnson W.O., Gardner I.A. and Singh R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. Applied Statistics, **52**, 63–76.

(i)   Test the hypothesis that the probability of a positive test result is the same for the two tests.

(9)

(ii)   Obtain approximate 95% confidence intervals for the proportion of positive test results for

(a)   the MAT,

(b)   the ELISA.

The researchers report that the true population proportion of positive blood samples is believed to be approximately 0.069. State, giving brief reasons, whether either or both of the approximate 95% confidence intervals you have calculated is consistent with this value.

(11)

5. (i) State the assumptions on which an independent (unpaired) two-sample t test is based.

(4)

An individual is considering purchasing a two-bedroomed terraced house in one of two adjacent towns in the north of England. To compare prices, he extracts some price data from one week's issue of the local newspaper's "Property Supplement" for all such properties advertised by estate agents with branches in both towns. He wishes to determine whether the mean price in one town differs from that in the other. The data extracted are given in the table below (units are thousands of pounds).

Town 1   77.50 74.95 74.50 60.00 45.00 25.00 25.00

Town 2   72.95 72.95 65.00 62.50 56.95 54.95 52.95
                49.95 46.95 35.00 34.95 30.00 29.95 25.00

(ii) Assuming that all the necessary assumptions hold, perform an independent two-sample t test and draw your conclusion.

(12)

(iii) Town 1 has a considerably larger population, and greater numbers of all types of properties, than town 2. Taking note of this information, and of the way in which the data were obtained, discuss critically whether there is a valid basis for the test in part (ii).

(4)

Turn over

6. (i) Explain in what circumstances the Mann-Whitney U test might be preferred, rather than the t test, when comparing two independent samples.

(4)

The lifetimes of two electronic components, A and B, are to be compared by a manufacturer of televisions with the intention of using the type with the longer average lifetime. The manufacturer samples 10 components of each type at random from large batches of them and, in controlled conditions, tests the length of time to failure (the lifetime), resulting in the data (in days) given in the table below.

Component A	Component B
410	460
416	233
456	301
407	285
421	301
491	343
532	400
432	231
634	249
481	328

(ii) Draw a dot-plot for each sample. With reference to your answer in part (i), suggest which test might be preferred in this case.

(6)

(iii) Perform a Mann-Whitney U test on the data given in the above table. Briefly explain your conclusion in a manner appropriate for the television manufacturer to understand.

(10)

7

Turn over

7.   (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.

(5)

(ii)   A psychology researcher has the hypothesis that effective use of leisure time helps to reduce stress. In particular, she suggests that play activity is most effective when the subject feels it is free play, not directed by others.

The researcher recruited 36 college students and divided them randomly into three groups. One group received highly controlled play experience, one received a low level of control and one group performed what they would see as work rather than play.

All subjects first performed a 30-minute stress-producing task, working through mathematics problems while hearing periodic bursts of loud noise through headphones.

Next, each subject had 10 minutes at one of the three play activities described above, "high" or "low" control or "work".

Finally, the subjects attempted to solve two geometric puzzles, one of which was insoluble – but they were not told this. Persistence on the insoluble puzzle (measured in time in total seconds spent on the puzzle before giving up) was the response variable measured and used to assess the effectiveness of the play period in reducing the stress created by the work task. The table below gives the results.

High	Low	Work
347	504	398
567	420	492
424	583	97
239	183	357
256	279	184
682	381	554
435	118	354
666	317	275
825	359	198
102	77	163
601	336	284
384	197	155

Perform a one-way analysis of variance on these data and, by computing least significant differences, or otherwise, investigate differences between the three means.

Write a brief report for the researcher to use when interpreting the results.

(15)

8

Turn over

8.   In a survey of British adults (those aged over 16), there were questions about alcohol consumption. One of these questions related to their average weekly alcohol consumption. The table below shows trends in the distribution of average weekly alcohol consumption during the period 1992–1998.

Average Weekly alcohol consumption by sex and age: 1992 to 1998
Persons aged 16 and over

Age	1992	1994	1996	1998
	Mean number of units per week			
Men				
16–24	19.1	17.4	20.3	23.6
25–44	18.2	17.5	17.6	16.5
45–64	15.6	15.5	15.6	17.3
65 and over	9.7	10.0	11.0	10.7
Total	15.9	15.4	16.0	16.4
Women				
16–24	7.3	7.7	9.5	10.6
25–44	6.3	6.2	7.2	7.1
45–64	5.3	5.3	5.9	6.4
65 and over	2.7	3.2	3.5	3.3
Total	5.4	5.4	6.3	6.4
All persons				
16–24	12.9	12.3	14.7	16.6
25–44	11.8	11.4	11.9	11.4
45–64	10.2	10.2	10.5	11.6
65 and over	5.6	6.0	6.8	6.5
Total	10.2	10.0	10.7	11.0

(i)   Using the statistics given in the table, draw suitable diagrams to illustrate

   (a)   the overall trend in alcohol consumption amongst men and women during the period 1992–1998,

   (b)   similarities and differences between the 1998 age-specific alcohol consumption for males and females.

(10)

(ii)   Write a short report, suitable for publication in a serious newspaper, based on the diagrams you have produced in part (i) and any other aspects of the data you consider relevant.

(10)