EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2007

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 $\binom{n}{r}$ *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. An experiment was carried out to study the effect of two factors on the quality of pancakes. The two factors were percentage of whey used and whether or not a cereal additive was used. There were four levels of whey, giving $4 \times 2 = 8$ treatment combinations. Three pancakes were baked using each treatment combination in random order. Each pancake was rated in random order by an expert, the higher the quality rating, the better the pancake. The rating for each of the 24 pancakes is shown below.

	Amount of whey			
	0%	10%	20%	30%
No additive	4.4 4.5 4.3	4.6 4.5 4.8	4.5 4.8 4.8	4.6 4.7 5.1
Additive	3.3 3.2 3.1	3.8 3.7 3.6	5.0 5.3 4.8	5.4 5.6 5.3

(i) Copy and complete the following analysis of variance table, perform the F tests and give your conclusions.

```
Analysis of Variance for rating

Source          DF         SS      MS   F ratio

Additive         *          *       *         *
Whey             *     6.6913       *         *
Additive*Whey    *     3.7246       *         *
Residual         *     0.4800       *

Total            *   11.4063
```

State the assumptions required for the F tests to be valid.

(9)

(ii) Expand on your conclusions in (i) by constructing a suitable graph, and comment on the graph.

(7)

(iii) Given the analysis so far, comment on how well the model explains the variation in the data. Describe any further checks you would make on the appropriateness of the model.

(3)

(iv) Give one question you would ask about how the experiment was conducted.

(1)

Turn over

2. A study was carried out to determine the effect of the presence or absence of a company safety programme on the number of work hours lost due to work-related accidents. Fifty companies were selected from a particular business sector that contains a large number of companies. A lost work hours ratio was calculated for each company by dividing the total number of hours lost over a given period by the number of employees. The data (given for convenience in order of size) are shown below.

Companies with no safety programme (sample A)

0.0083	0.0089	0.0091	0.0094	0.0116	0.0133	0.0133	0.0145
0.0153	0.0166	0.0168	0.0169	0.0169	0.0173	0.0179	0.0183
0.0186	0.0189	0.0195	0.0202	0.0204	0.0211	0.0217	0.0223
0.0223	0.0230	0.0233	0.0234				

Companies with a safety programme (sample B)

0.0029	0.0070	0.0072	0.0085	0.0092	0.0095	0.0099	0.0106
0.0106	0.0111	0.0123	0.0128	0.0138	0.0142	0.0143	0.0153
0.0155	0.0164	0.0189	0.0198	0.0213	0.0245		

(i) Use the summary data given below to carry out a two-sample *t* test to determine whether there is evidence that a safety programme reduces the lost work hours ratio. State carefully the null and alternative hypotheses.

	Number of observations	*Sample mean*	*Sample standard deviation*
Sample A	28	0.01711	0.00462
Sample B	22	0.01298	0.00513

(7)

(ii) State the assumptions required for validity of the test in (i), and draw a picture of the data that provides a suitable "by eye" assessment of whether or not the assumptions are valid. Discuss briefly what your picture shows.

(9)

(iii) In a situation where the assumptions, either individually or together, for a two-sample *t* test are not valid, give brief details of an alternative test that might be carried out.

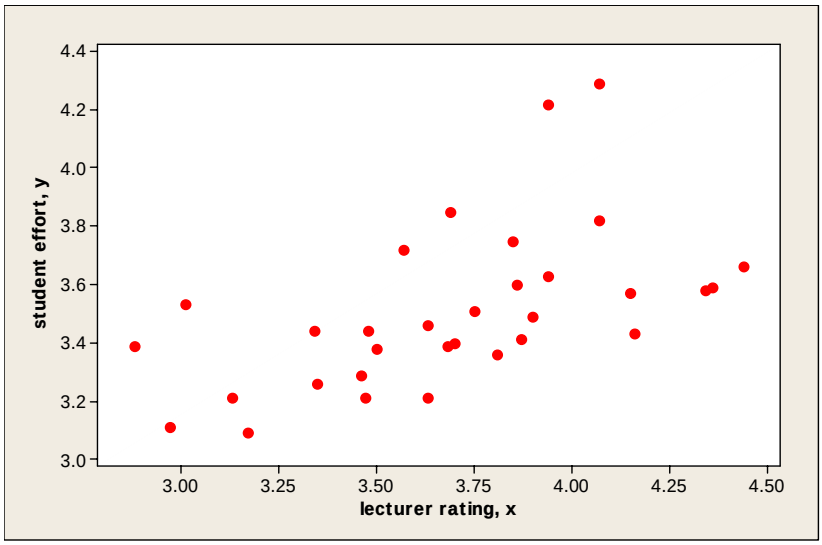
(4)

Turn over

3. A large number of first-year students at a university completed feedback questionnaires about each module attended. The students were asked to answer several questions, using a scale of 1 (low) to 5 (high). In particular, they assessed the quality of the module lecturer, and indicated the extent to which they had worked on the module to the best of their ability.

Data were collected from 32 modules; for module i, x_i was the average lecturer quality rating and y_i was the average student effort rating.

The scatter plot below illustrates the collected data. The question of interest is how student effort on a module is related to the rating of the lecturer.



The model $y_i = \alpha + \beta x_i + e_i$ is proposed, where the $\{e_i\}$ are independent random variables each distributed as $N(0, \sigma^2)$.

(i) Comment on the suitability of the model, based on the scatter plot, and comment on the choice of student effort as the response variable. (4)

(ii) Find the values taken by the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β, given the following summary information.

$$\sum_{i=1}^{32} x_i = 118.17 \quad \sum_{i=1}^{32} y_i = 112.29 \quad \sum_{i=1}^{32} x_i y_i = 416.52 \quad \sum_{i=1}^{32} x_i^2 = 441.47$$

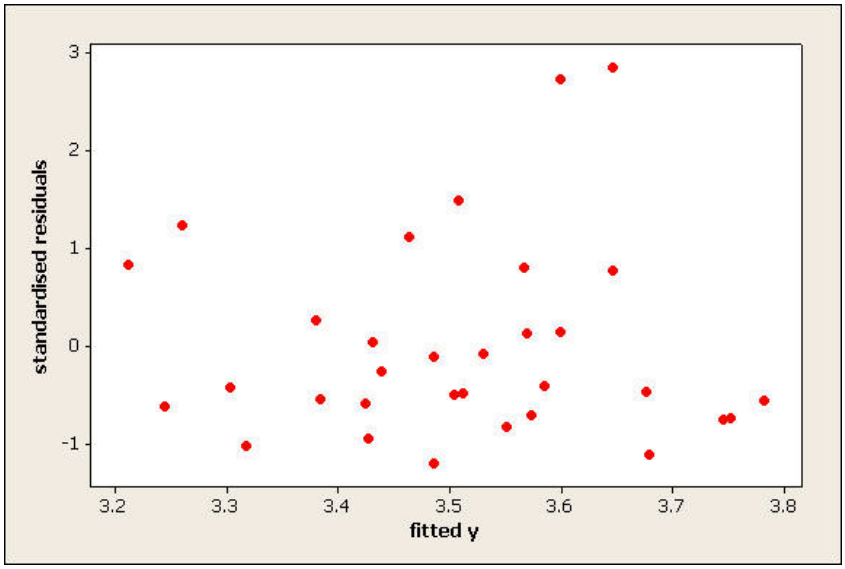
(5)

(iii) The standard errors of $\hat{\alpha}$ and $\hat{\beta}$ are respectively 0.3844 and 0.1035. Construct 95% confidence intervals for α and for β. (3)

(iv) Describe the practical interpretation of β and comment on the relationship between student effort and lecturer rating. (4)

(v) A plot of the standardised residuals against the fitted y-values is shown **on the next page**. Comment on how well this supports the assumptions made in the model. (4)

The plot of standardised residuals is on the next page



Turn over

4. (a) An insurance company has three types of insurance policy. Four sales staff sell these policies. The table below shows the number of policies of each type sold by each of the sales staff in a given time period. Analyse the data to see whether or not it is reasonable to assume that the type of policy sold is independent of which member of the sales team is doing the selling.

		Sales person			
		1	*2*	*3*	*4*
Policy type	*A*	100	90	80	70
	B	120	110	115	110
	C	130	140	145	150

(12)

(b) An insurance company provides buildings insurance cover for 200 properties in area 1 and 300 properties in area 2. Over a period of one year, eight of the properties in area 1 and five of the properties in area 2 give rise to a claim. The chance of any property giving rise to more than one claim in any given year is so small that it can be neglected. It can also be assumed that claims are independent of each other.

Carry out a significance test for the hypothesis that the difference between the annual claim rates for the two areas is zero, and state your conclusions

(8)

Turn over

5. The number of telephone enquiries received by an operative at a call centre during each minute is modelled by a Poisson distribution with mean μ, and the numbers of enquiries received in non-overlapping intervals are independent.

(i) The numbers of enquiries received in n one-minute intervals are represented by the observations $x_1, x_2, \ldots, x_n$. Write down the likelihood function in terms of μ and hence find the maximum likelihood estimator, $\hat{\mu}$, of μ. Find an estimator of the variance of $\hat{\mu}$.

(7)

(ii) The numbers of enquiries received by an operative over a period of 150 non-overlapping one-minute intervals are shown in the table below.

Calculate the mean number of enquiries received per minute.

Carry out an appropriate statistical test to determine whether or not a Poisson distribution is a suitable model for these data.

Number of enquiries per minute	*Frequency*
0	39
1	66
2	29
3	10
4	4
5	2
6 or more	0

(10)

(iii) Construct an approximate 95% confidence interval for the mean μ.

(3)

6. The following table shows the quarterly sales (in £100,000 and adjusted for inflation) of a certain soft drink for the years 2003 to 2006.

Year	*1st quarter*	*2nd quarter*	*3rd quarter*	*4th quarter*
2003	2	4	22	14
2004	9	11	30	21
2005	16	18	38	29
2006	24	26	47	36

(i) Plot the data and describe the main features.

(6)

(ii) Write down a suitable additive model for the sales, Y_t, briefly explaining the meaning of the terms in the model.

(2)

(iii) From the table below, describe the moving average formula used and explain why it is a suitable choice for these data. Add the trend line to your plot in (i). Calculate and plot the residuals and comment on whether the model has adequately described the data.

sales	MA(trend)	detrended
2		
4		
22	11.375	10.625
14	13.125	0.875
9	15.000	-6.000
11	16.875	-5.875
30	18.625	11.375
21	20.375	0.625
16	22.250	-6.250
18	24.250	-6.250
38	26.250	11.750
29	28.250	0.750
24	30.375	-6.375
26	32.375	-6.375
47		
36		

(9)

(iv) Describe briefly a method for estimating the value of total sales in the year 2007. State any assumptions you need to make.

(3)

7. In the UK National Lottery, 49 balls (numbered 1 through to 49) are used. In each draw, 6 of the balls are selected supposedly at random. After eleven years of the running of the UK National Lottery, a national newspaper published a list of how many times each of the 49 numbers had been selected over that time. The article cited ball number 20 as the unluckiest as it had been drawn the smallest number of times, and ball number 38 as the luckiest as it had been drawn the largest number of times. The data are shown below.

No.	*Frequency*	*No.*	*Frequency*	*No.*	*Frequency*	*No.*	*Frequency*	*No.*	*Frequency*
1	126	11	138	21	112	31	137	41	104
2	128	12	131	22	124	32	136	42	124
3	124	13	108	23	143	33	133	43	150
4	118	14	120	24	117	34	120	44	149
5	121	15	117	25	147	35	131	45	132
6	129	16	107	26	127	36	118	46	125
7	131	17	120	27	130	37	114	47	144
8	119	18	126	28	128	38	156	48	139
9	126	19	130	29	125	39	115	49	124
10	132	20	101	30	136	40	136		

The article suggested that picking certain numbers gave a better chance of winning a prize.

(i) Describe how you would plot the data in a way that represents the data accurately but puts the variation in the frequencies into perspective. Illustrate this using numbers 1 to 6.

(6)

(ii) Describe in simple terms how you would test for evidence that some balls are luckier or unluckier than others. Express your explanation in a way that could be understood by someone with no statistical training, for example as a short article that might be suitable for a newspaper.

[**Note.** When the hypothesis that all numbers are equally likely to be selected was tested, using the data above, the result was not significant at the 5% level.]

(10)

(iii) If your audience remained unconvinced by your argument, what further data might you consider in order to help to demonstrate the likely variability in the frequencies?

(4)

8. (i) Describe the effects of *sampling with replacement* and *sampling without replacement* in the context of drawing a random sample from (a) a modest sized population, (b) a large population.

(5)

(ii) A consumer organisation wishes to investigate a particular aspect of household expenditure for households in England. It is decided that a random sample of households will be drawn from each of a random sample of geographical areas across England. The country is already divided into administrative areas and these are felt to form a suitable first stage sampling frame.

The following table lists all the administrative areas in England.

	Area		*Area*		*Area*		*Area*
1	Avon	13	Dorset	25	Lancashire	37	Somerset
2	Bedfordshire	14	East Sussex	26	Leicestershire	38	South Yorkshire
3	Berkshire	15	Essex	27	Lincolnshire	39	Staffordshire
4	Buckinghamshire	16	Gloucestershire	28	Merseyside	40	Suffolk
5	Cambridgeshire	17	Greater Manchester	29	Norfolk	41	Surrey
6	Cheshire	18	Hampshire	30	North Yorkshire	42	Tyne & Wear
7	Cleveland	19	Herefordshire	31	Northamptonshire	43	Warwickshire
8	County Durham	20	Hertfordshire	32	Northumberland	44	West Midlands
9	Cornwall	21	Humberside	33	Nottinghamshire	45	West Sussex
10	Cumbria	22	Inner London	34	Outer London	46	West Yorkshire
11	Derbyshire	23	Isle of Wight	35	Oxfordshire	47	Wiltshire
12	Devon	24	Kent	36	Shropshire	48	Worcestershire

Use the first column of double random digits in the table on page 18 of the "Statistical tables for use in examinations" to select a random sample of six areas from the above list.

(6)

(iii) Describe two other types of sampling method that might be used here. If the eventual aim is to obtain a representative sample of households, discuss the relative merits of the two methods.

(9)