EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2004

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. (i) By means of examples, or otherwise, explain how you might decide when to perform the following significance tests. State the null hypothesis under test in each case.

(a) Two independent samples t test.

(b) Matched pairs t test.

(c) Mann-Whitney test (also known as the Wilcoxon rank sum test).

(d) Wilcoxon signed-rank test.

(12)

(ii) A firm uses aptitude tests A and B as part of the procedure of deciding which applicants to interview. Each applicant takes one of the two tests (allocated at random) and receives a score.

The scores on test A obtained by seven applicants for a particular job were

52, 61, 68, 50, 60, 58, 64.

The scores on test B obtained by seven other applicants for the job were

62, 67, 69, 73, 72, 59, 71.

The personnel manager wishes to know whether the two aptitude tests give similar average scores. Formulate appropriate null and alternative hypotheses. Explain briefly whether you would use a parametric or a non-parametric test here, and why. Carry out an appropriate test, and report your conclusion.

(8)

2. A study is undertaken to compare how the time taken by a particular computer software application to read in data files varies with the format of the file. Random samples of eight files in each of three formats were taken. The times in milliseconds to read the files in each format are shown below.

Standard format	*First alternative format*	*Second alternative format*
2.02	2.18	1.87
1.99	1.84	2.56
2.01	1.99	2.02
1.88	1.91	2.44
2.27	2.09	2.46
2.36	2.08	2.61
2.31	2.23	2.45
2.35	1.84	2.11

(i) Specify two questions you would want to ask the collector of the data before deciding what sort of analysis might be appropriate.

(2)

(ii) State the model for a one-way analysis of variance and the assumptions required for its validity. How might you investigate whether the assumptions are satisfied?

(5)

(iii) Perform a one-way analysis of variance on the above data, stating the null hypothesis, and interpret the result.

(7)

(iv) Test the null hypothesis that the population mean times of the two possible alternative formats are the same against an alternative hypothesis to be stated. Obtain a 95% confidence interval for the difference between the mean times of these two possible alternative formats.

(6)

3. A drug which was thought to affect the plasma cholesterol level of humans was tested in two experiments by comparing it with a placebo. (A placebo is a chemically inert substance which is known to have no physical effect but which is similar in appearance to a conventional medicine.) In experiment 1, 10 subjects chosen at random were given the drug on one day and the placebo a week later, by which time it was thought that any effect of the drug would have worn off. Experiment 2 used 20 different subjects, also chosen at random. Of these, ten were randomly allocated to the drug and the other ten were given the placebo at the same time. Blood samples were taken from all subjects two hours after they had taken the drug or placebo, and the cholesterol levels were obtained.

The table below shows the cholesterol levels of the 40 blood samples in mg/100ml.

Experiment 1			**Experiment 2**			
	Responses			**Responses**		**Responses**
Subject	*Drug*	*Placebo*	*Subject*	*Drug*	*Subject*	*Placebo*
A	196	192	K	203	U	168
B	190	187	L	197	V	174
C	155	149	M	210	W	205
D	199	200	N	153	X	251
E	190	183	O	197	Y	160
F	203	203	P	225	Z	173
G	237	242	Q	157	AA	162
H	202	194	R	236	BB	179
I	228	223	S	171	CC	202
J	212	207	T	222	DD	199

(i) Which of these two do you think is the better experimental design, and why?

(3)

(ii) Calculate a 95 per cent confidence interval for the difference between the mean cholesterol levels of subjects on drug and placebo using the results of experiment 1, stating any assumptions that you make.

(7)

(iii) Calculate a 95 per cent confidence interval for the difference between the mean cholesterol levels of subjects on drug and placebo from the results of experiment 2, stating any assumptions that you make.

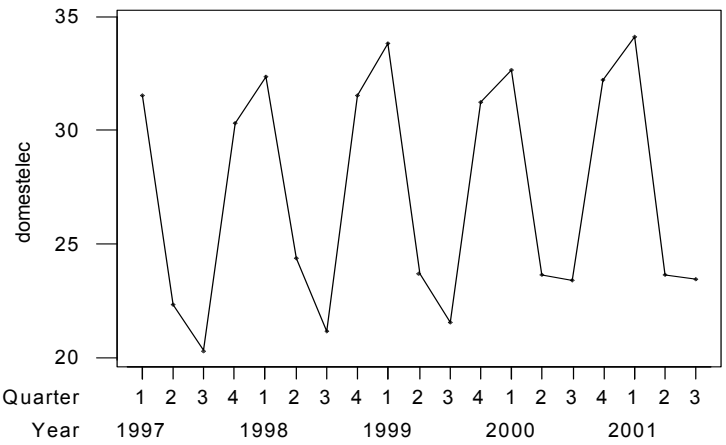
(7)

(iv) Does the drug appear to increase cholesterol level? Justify your answer.

(3)

4. The display and plot below show quarterly domestic sales (Sales) in terawatt hours by the public electricity supply system in the UK from the first quarter of 1997 to the third quarter of 2001. The display also shows centred 4-quarterly moving averages (MA), and the values of the differences "Sales – MA".

Year	*Quarter*	*Sales*	*MA*	*Sales – MA*
1997	1	31.54	*	*
1997	2	22.33	*	*
1997	3	20.29	26.22	–5.93
1997	4	30.30	26.57	3.73
1998	1	32.35	26.93	5.42
1998	2	24.36	27.20	–2.84
1998	3	21.16	27.54	–6.38
1998	4	31.54	27.64	3.90
1999	1	33.85	27.61	6.24
1999	2	23.69	27.62	–3.93
1999	3	21.55	27.43	–5.88
1999	4	31.22	27.27	3.95
2000	1	32.64	27.49	5.15
2000	2	23.64	27.84	–4.20
2000	3	23.37	28.15	–4.78
2000	4	32.20	28.32	3.88
2001	1	34.10	28.33	5.77
2001	2	23.61	*	*
2001	3	23.46	*	*



(i) Give the calculation leading to the value of 26.22 for the MA for 1997 quarter 3. (4)

(ii) Using the differences given for "Sales – MA", estimate the pattern of the seasonal variation in sales. (8)

(iii) Why is it good practice in the course of making such estimates to plot the data and the moving average trend as time series? (4)

(iv) Using the estimates found in part (ii), correct the 2000 Sales figures for seasonal fluctuations.

Could you similarly correct the 2001 figures? Explain your answer.

(4)

Turn over

5. The table shows, for South Africa, details of the numbers of road traffic collisions and casualties associated with them, in four years.

(i) Examine the changes over time, considering in particular whether or not accidents are becoming less serious. Support your answer by any diagrams and calculations you think are appropriate.

(14)

(ii) What other background information might be useful in interpreting the figures in this table, and why?

(6)

	Numbers			
	1994	*1995*	*1996*	*1997*
Collisions				
Fatal	8140	8335	7850	7790
Major	22594	23988	22707	23059
Minor	60199	61716	54190	57391
No injury	377064	406194	436027	417748
Total	**467997**	**500233**	**520774**	**505988**
Casualties				
Killed	9981	10256	9848	9691
Seriously injured	36548	39780	38473	39302
Slightly injured	91887	96689	86728	91760
Total	**138416**	**146725**	**135049**	**140753**

Source: Statistics South Africa.

6. Some measurements of the volume v (cc) of a casting at seven different temperatures t (in degrees Celsius) of the casting are shown below.

Temperature	18	19	20	21	22	23	24
Volume	10.10	10.80	11.45	12.18	12.80	13.35	15.90

$\Sigma t^2 = 3115$ $\Sigma v^2 = 1092.9774$ $\Sigma tv = 1842.03$

Two linear regressions have been estimated, one using all seven pairs of measurements, and one omitting the last measurement listed (temperature = 24°C). Edited results are shown below; some quantities relating to the first regression have been omitted deliberately.

Regression Analysis: volume versus temperature – all measurements

```
The regression equation is volume = ** + ** temperature

Predictor        Coef     SE Coef        T        P
Constant          **      2.386      -2.31    0.069
temperature       **      0.1131      7.53    0.001

S = 0.5986      R-Sq = **%
```

Regression Analysis: volume versus temperature – omitting measurement at 24^0C

```
The regression equation is volume = -1.68 + 0.657 temperature

Predictor        Coef       SE Coef         T        P
Constant       -1.6797     0.2803      -5.99    0.004
temperature     0.65657    0.01362     48.19    0.000

S = 0.05700     R-Sq = 99.8%
```

(i) Plot the data and comment.

(4)

(ii) Calculate the missing estimated coefficients in the first regression, and find the coefficient of determination R^2.

(5)

(iii) Which regression do you think is better, and why?

(4)

(iv) For your preferred regression, interpret its estimated coefficients and its R^2. Explain the meanings of the values of "SE Coef", "T" and "P" given above for it. Why are these values useful?

(7)

Turn over

7. A random sample of 250 people has been taken from the 1985 US Current Population Survey. Some computer output relating to the variables WAGE (wage in dollars per hour) and SECTOR (0 = Other, 1 = Manufacturing, 2 = Construction) is given below.

(i) "TrMean" in the output is the trimmed mean. This is the mean obtained when the largest 5% and smallest 5% (rounded to the nearest integer) of values are removed. What advantage might there be in using a trimmed mean rather than a conventional mean? (4)

(ii) Draw a boxplot of WAGE for each of the three sectors. Indicate on your boxplots any values which might be regarded as outliers. (6)

(iii) Write a report comparing wages by sector, based on the computer output and your answer to part (ii). (10)

Descriptive Statistics: wages by sector

Variable	SECTOR	N	Mean	Median	TrMean	StDev
WAGE	0	188	9.609	8.500	9.055	5.920
	1	48	9.492	9.245	9.293	4.244
	2	14	9.507	10.375	9.529	3.633

Variable	SECTOR	SE Mean	Minimum	Maximum	Q1	Q3
WAGE	0	0.432	1.750	44.500	5.250	12.120
	1	0.613	3.000	20.400	5.890	11.658
	2	0.971	3.750	15.000	6.500	11.808

Data Display – Wages in sample from sector 0 (in ascending order)

1.75	3.35	3.35	3.40	3.45	3.50	3.50	3.50	3.50
3.50	3.51	3.55	3.56	3.65	3.65	3.75	3.75	3.80
3.84	4.00	4.00	4.13	4.17	4.25	4.25	4.28	4.35
4.35	4.50	4.50	4.50	4.50	4.55	4.55	4.55	4.85
5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.13	5.20
5.21	5.25	5.25	5.30	5.50	5.50	5.50	5.50	5.71
5.71	5.75	5.75	5.80	5.80	5.85	5.87	6.00	6.00
6.25	6.25	6.25	6.25	6.25	6.25	6.50	6.67	6.75
6.85	7.00	7.00	7.00	7.00	7.14	7.50	7.50	7.50
7.53	7.67	7.69	7.75	7.80	7.88	7.96	8.00	8.00
8.20	8.49	8.50	8.50	8.50	8.56	8.63	8.75	8.75
8.75	8.75	8.75	8.80	8.85	8.89	8.90	8.93	9.00
9.00	9.00	9.00	9.00	9.15	9.33	9.37	9.42	9.50
9.50	9.60	9.60	9.86	10.00	10.00	10.00	10.00	10.00
10.00	10.00	10.20	10.25	10.50	10.81	11.11	11.25	11.25
11.35	11.79	11.84	12.00	12.00	12.00	12.16	12.50	12.50
12.65	12.67	13.00	13.12	13.16	13.20	13.33	13.45	13.45
13.45	13.65	13.95	14.00	14.29	14.53	14.67	15.00	15.00
15.00	15.56	15.79	15.95	16.00	16.14	16.65	17.25	18.00
19.98	19.98	20.00	20.00	20.50	20.55	22.20	22.20	22.50
22.50	22.50	24.98	24.98	24.98	24.98	26.00	44.50	

Data Display – Wages in sample from sector 1 (in ascending order)

3.00	3.35	4.00	4.50	4.62	4.80	4.85	4.95	5.10
5.40	5.65	5.77	6.25	6.50	6.67	6.75	6.80	7.00
7.78	8.40	8.50	8.89	9.00	9.24	9.25	10.00	10.00
10.50	10.53	10.58	10.62	11.00	11.00	11.25	11.32	11.50
11.71	12.00	12.00	12.00	12.50	13.89	15.00	15.38	16.42
19.00	19.98	20.40						

Data Display – Wages in sample from sector 2 (in ascending order)

3.75	4.30	5.00	7.00	7.30	8.90	10.00	10.75	10.78
11.43	11.67	12.22	15.00	15.00				

8. The table below shows, for graduate entrants to an organisation during the last five years, the numbers of males and the numbers of females with each class of degree. You may treat the data as if they were a random sample from a large population.

Class of degree	*Male*	*Female*
1st	18	17
2nd	90	50
3rd	12	18
Pass	61	32

(i) Test whether there is an association between sex and class of degree. Write a short report to the personnel manager of the organisation describing your findings in non-technical language.

(9)

(ii) The national percentage of all graduates obtaining first class degrees during the last five years was 8.3. Assuming the number of graduate entrants having first class degrees follows a binomial distribution, investigate whether the overall proportion of graduates with first class degrees recruited by this organisation in this period is more than the average. Why might the assumption of a binomial distribution not be strictly valid?

(6)

(iii) Use the summary data from the last five years to investigate whether the population proportion of graduate entrants to this organisation who have second class degrees is the same for males as for females.

(5)