EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2002

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.*

Note that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 11 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 data are from Altman (1991), *Practical Statistics for Medical Research*, and represent measurements of foetal head circumference (cm) made by four observers. Each observer made three independent measurements of the head circumference for each foetus.

Foetus	Observer			
	1	2	3	4
	14.3	13.6	13.9	13.8
1	14.0	13.6	13.7	14.7
	14.8	13.8	13.8	13.9
	19.7	19.8	19.5	19.8
2	19.9	19.3	19.8	19.6
	19.8	19.8	19.5	19.8
	13.0	12.4	12.8	13.0
3	12.6	12.8	12.7	12.9
	12.9	12.5	12.5	13.8

The analysis of variance (ANOVA) table is

Source	DF	SS	MS	F
Foetuses	2	****	****	****
Observers	****	****	****	****
Foetuses × Observers	****	0.562	****	****
Error	****	****	0.0767	
Total	35	327.610		

(i) Complete the ANOVA table and use it to assess what evidence the experiment provides regarding the main effects and the interaction between the factors foetus and observer.

(10)

(ii) Draw a simple diagram using the twelve mean values which illustrates the effects of foetus, observer and their interaction.

(5)

(iii) Summarise your conclusions in non-technical language which the experimenter would understand.

(5)

2. (a) As part of an investigation of student expenditure patterns at a UK university, a random sample of students was taken. It was found that a high proportion of them had mobile telephones. The following summary data, in £ per week, refer to the expenditure on mobile telephone calls of all students. (Note that the sample includes students who did not possess mobile telephones.)

	Sample size	*Sum of expenditures*	*Sum of squares of expenditures*
Males	87	1098.60	32234.71
Females	63	887.75	25810.04

(i) Test whether there is any difference between the mean expenditures of males and females, stating your null and alternative hypotheses clearly.

(5)

(ii) On what assumptions, if any, is your test based? Comment on whether you feel the test result to be reliable, and on any further information which should be obtained in order to throw greater light on the impact of mobile telephones on student expenditure.

(5)

Part (b) of question 2 is on the next page

This is part (b) of question 2

(b) The table below shows, for each of 15 companies randomly selected from the largest 1000 companies in a particular country, the dividends announced in the years 1999 (x_1) and 2000 (x_2), together with the difference x ($x = x_2 - x_1$). All figures for 1999 and 2000 are expressed as percentages of the company's dividend in 1995.

Company	x_1 (1999)	x_2 (2000)	x
A	104.8	106.1	1.3
B	106.2	107.7	1.5
C	108.7	111.3	2.6
D	108.5	108.9	0.4
E	105.0	106.3	1.3
F	105.1	106.6	1.5
G	122.1	125.2	3.1
H	112.9	117.4	4.5
I	111.3	113.8	2.5
J	106.7	108.8	2.1
K	107.8	110.4	2.6
L	110.8	113.3	2.5
M	110.7	113.5	2.8
N	105.4	106.9	1.5
O	110.4	112.8	2.4

Note: $\Sigma x_1 = 1636.4$, $\Sigma x_2 = 1669.0$, $\Sigma x = 32.6$,
$\Sigma x_1^2 = 178797.12$, $\Sigma x_2^2 = 186071.08$, $\Sigma x^2 = 84.18$.

(i) Carry out a test based on the differences between the 1999 and 2000 figures in which you take account of the fact that each pair of observations refers to one company in two successive years.

(5)

(ii) A commentator reported on the data by comparing the difference between the sample means for 1999 and 2000 with the standard error calculated as

$$\sqrt{\frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{28}},$$

giving a value for t of 1.241. What conclusion would the commentator reach?

(2)

(iii) Explain and discuss any difference between the commentator's conclusion and that from your test in part (i).

(3)

3. The table below shows the UK Gross Domestic Product (GDP, y) for the years 1989 – 1999, with years also coded as t = year – 1994. The figures are given in units of £bn, and are expressed in 1995 prices.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
t	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	655.2	659.5	649.8	650.3	665.4	694.6	714.0	732.2	757.9	777.9	794.4

Source: United Kingdom National Accounts, 2000 edn, table 1.1.

Note: $\Sigma y = 7751.2$, $\Sigma y^2 = 5491108.76$, $\Sigma t = 0$, $\Sigma t^2 = 110$, $\Sigma ty = 1706.3$.

(i) A model $y = \alpha + \beta t + \varepsilon$, where ε is a random error term with the usual properties, is proposed for the data. Obtain least squares estimates of α and β, and calculate r^2 (the coefficient of determination). Also estimate σ^2, the variance of ε, and obtain estimates of the standard errors of the coefficients α and β.

(6)

(ii) What are "the usual properties" of the errors? How realistic is the assumption that the errors have these properties? (You are not expected to describe or conduct any tests.)

(2)

(iii) Interpret the value of r^2, and the values of your estimates of α and β.

(3)

(iv) Draw a time chart of the data and superimpose your estimated function on it.

(5)

(v) Use your estimated model to predict GDP in 2000 and 2010, and comment on your predictions in the light of your graph.

(4)

Turn over

4. Quarterly data relating to passenger traffic on United Kingdom railways, published in the *Monthly Digest of Statistics* August 1999 and August 2001, are fitted by a centred four-quarter arithmetic average. The differences between the actual data and the moving average are obtained. The results are as follows.

		National Rail, passenger kilometres, millions		
Period		*Actual*	*Trend*	*Difference*
1996	Q1	7554	na	na
	Q2	7817	na	na
	Q3	8101	8005.500	95.500
	Q4	8330	8125.625	204.375
1997	Q1	7994	8277.500	–283.500
	Q2	8338	8443.875	–105.875
	Q3	8795	8594.125	200.875
	Q4	8967	8713.375	253.625
1998	Q1	8559	8801.500	–242.500
	Q2	8727	8895.125	–168.125
	Q3	9111	9009.500	101.500
	Q4	9400	9134.875	265.125
1999	Q1	9041	9277.500	–236.500
	Q2	9248	9397.750	–149.750
	Q3	9731	9512.375	218.625
	Q4	9742	9664.625	77.735
2000	Q1	9616	9885.750	–269.750
	Q2	9891	9969.500	–78.500
	Q3	10857	na	na
	Q4	9286	na	na

Note: "na" means "not available"

(i) Using the method of differences from a moving arithmetic average, estimate the seasonal pattern in the observed data and hence correct that series for seasonal fluctuations.

(12)

(ii) Comment on the results.

(4)

(iii) Describe another method of seasonal correction which might have been preferable. What advantages would it have had in this instance?

(4)

5. Data were collected from the *Financial Times* of 11 October 2001 on the dividend yields of 36 Bank shares (Banks), 46 Electronic and Electrical Equipment company shares (E&EEq) and 88 Support Services company shares (SS). (Support Services companies provide services such as catering and security to manufacturing and other companies.) These observations were read into a computer using the Minitab package and arranged in ascending order as shown in the edited output **on this page and the next**.

Note that the command `DESCRIBE` is used to obtain the mean, median, quartiles, etc, of the three sets of observations as shown, and the commands `DOTPLOT` and `STEM-AND-LEAF` generate appropriate diagrams.

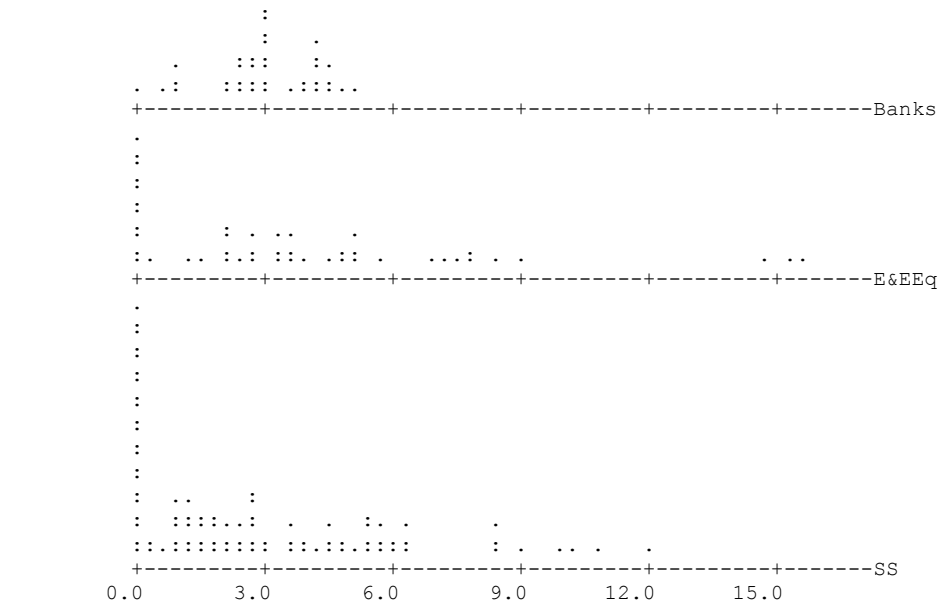
(i) Draw boxplots of each of the three sets of observations, indicating any outliers. (6)

(ii) Explain how stem and leaf diagrams display the observations, and how they relate to the dotplots. (4)

(iii) Write a report on the patterns of the three sets of yields, based on the output generated by the `DESCRIBE` command, your boxplots, and the computer-generated dotplots and stem and leaf diagrams. (10)

```
MTB > PRINT 'Banks'

    0.0    0.7    0.8    0.9    1.0    2.1    2.1    2.3    2.4    2.5
    2.5    2.7    2.7    2.8    2.8    2.9    3.0    3.0    3.0    3.0
    3.1    3.1    3.1    3.7    3.8    3.8    4.1    4.1    4.1    4.2
    4.3    4.4    4.5    4.5    4.7    5.1

MTB > PRINT 'E&EEq'

    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
    0.0    0.4    1.3    1.5    2.0    2.0    2.1    2.1    2.4    2.6
    2.6    2.8    3.2    3.3    3.4    3.5    3.5    3.7    3.9    4.4
    4.8    4.9    5.0    5.1    5.2    5.7    6.9    7.3    7.6    7.8
    7.9    8.3    9.0   14.6   15.2   15.6

MTB > PRINT 'SS'

    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
    0.1    0.3    0.4    0.5    0.9    0.9    1.0    1.0    1.0    1.2
    1.3    1.3    1.3    1.3    1.4    1.4    1.5    1.6    1.7    1.8
    1.8    1.9    2.0    2.2    2.2    2.3    2.5    2.5    2.6    2.6
    2.6    2.7    2.8    2.8    2.9    3.0    3.5    3.6    3.7    3.8
    3.9    4.2    4.4    4.5    4.6    4.7    4.8    5.2    5.3    5.4
    5.5    5.5    5.6    5.7    5.7    5.9    6.0    6.2    6.3    6.4
    8.3    8.3    8.4    9.0    9.8   10.1   10.7   11.9


MTB > DESCRIBE 'Banks' 'E&EEq' 'SS'

Variable             N       Mean     Median       StDev    SE Mean
Banks               36      2.994      3.000       1.232      0.205
E&EEq               46      3.948      3.250       3.982      0.587
SS                  88      2.934      2.200       2.909      0.310

Variable       Minimum    Maximum         Q1         Q3
Banks            0.000      5.100      2.425      4.100
E&EEq            0.000     15.600      0.300      5.325
SS               0.000     11.900      0.325      4.775
```

```
MTB > DotPlot 'Banks'-'SS';
SUBC>   Same.
```




```
MTB > Stem-and-Leaf 'Banks'.
   N  = 36   Leaf Unit = 0.10

    1    0 0
    4    0 789
    5    1 0
    5    1
    9    2 1134
   16    2 5577889
   (7)   3 0000111
   13    3 788
   10    4 111234
    4    4 557
    1    5 1

MTB > Stem-and-Leaf  'E&EEq'.
N  = 46    Leaf Unit = 0.10

   12    0 000000000004
   14    1 35
   22    2 00114668
   (7)   3 2345579
   17    4 489
   14    5 0127
   10    6 9
    9    7 3689
    5    8 3
    4    9 0
    3   10
    3   11
    3   12
    3   13
    3   14 6
    2   15 26

MTB > Stem-and-Leaf  'SS'.
N  = 88               Leaf Unit = 0.10

   26    0 00000000000000000000134599
   42    1 0002333344567889
  (13)   2 0223556667889
   33    3 056789
   27    4 245678
   21    5 234556779
   12    6 0234
    8    7
    8    8 334
    5    9 08
    3   10 17
    1   11 9
```


Turn over

6. A society has two grades of membership (Grade I and Grade II) and a worldwide membership of about 7000 individuals. About 20% of the members fall into Grade II, and about 75% of all members are concentrated into three geographical areas (A, B, C), the rest being spread throughout the rest of the world.

The society publishes a journal which is sent by post to all members, and it wishes to carry out a survey to discover if members find the journal useful, and what aspects of their subject they most wish to see covered in the journal. The society is particularly anxious to discover whether members of both grades find the journal useful.

The society keeps its membership list as a computer file, with one record for each member. The records are stored in alphabetical order of members' names, though the secretary is assured that separate lists could be provided for the different grades of membership.

Consider five possible methods for selecting a sample of members to receive, by post, a questionnaire for this purpose:

simple random sampling,
stratified random sampling,
quota sampling,
cluster sampling,
systematic sampling.

For each method, discuss

(i) whether it would be possible to use this method of sampling for this purpose,

(ii) whether the method would be a good one to choose for the purpose.

(4 marks for each method)

7. The time a cashier takes, in minutes, to serve a customer in a supermarket is modelled as a random variable X with probability density function

$$f(x) = \lambda^2 x e^{-\lambda x}, \quad x \geq 0,$$

where λ is an unknown parameter ($\lambda > 0$). It can be shown that the mean of this distribution is $2/\lambda$, and that the cumulative distribution function is given by

$$F(x) = 1 - (1 + \lambda x) e^{-\lambda x}, \quad x \geq 0.$$

(i) Find the value of x for which $f(x)$ takes its maximum value (the mode of X).

(4)

(ii) Draw a rough sketch of $f(x)$ against x for the case $\lambda = 1$.

(2)

(iii) Give a theoretical (that is, mathematical) explanation for the large difference between the mean and the mode of X. Do you consider that this distribution might be a good model for the times taken to serve customers in a supermarket?

(3)

(iv) A random sample of n customers is selected, and their service times $x_1, x_2, \ldots, x_n$ are measured. Find the maximum likelihood estimator and the method of moments estimator of λ. Comment on your findings.

(6)

(v) The manager sets a target that the mean service time should be 2 minutes; that is, $\lambda = 1$. To test this, a random sample of 200 service times was obtained. The data were then grouped into the frequency distribution shown below. The expected frequencies shown were calculated on the assumption that $\lambda = 1$. Calculate the remaining expected frequencies, and use the information to test the hypothesis that the data come from this distribution with $\lambda = 1$.

Range of x	*Observed frequency*	*Expected frequency*
$0.0 < x \leq 0.5$	21	18.04
$0.5 < x \leq 1.0$	30	34.81
$1.0 < x \leq 1.5$	36	
$1.5 < x \leq 2.0$	29	30.36
$2.0 < x \leq 2.5$	27	23.74
$2.5 < x \leq 3.0$	19	17.63
$3.0 < x \leq 3.5$	17	12.65
$3.5 < x \leq 4.0$	4	8.86
$4.0 < x \leq 4.5$	8	6.10
$4.5 < x \leq 5.0$	2	4.13
$x > 5.0$	7	

(5)

Turn over

8. A herbicide experiment was carried out on oats. This experiment was laid out in a completely randomised design and involved 4 herbicides with 10 plots per herbicide. The variable recorded (y) was the number of weeds seen on each plot. A review of the data suggested that the variances were not equal. The data were analysed both in their original form and after being transformed. Transformations tried were square root, logarithm and exponential. The following table contains values of the means and variances of the transformed data.

Herbicide	**Transformation**							
	None		$\sqrt{y}$		$\log_{10}(y)$		$\exp(y/100)$	
	mean	*variance*	*mean*	*variance*	*mean*	*variance*	*mean*	*variance*
A	99.30	477.42	9.903	1.3572	1.9857	0.01195	2.754	0.3047
B	79.20	143.76	8.877	0.4489	1.8943	0.00425	2.222	0.0726
C	139.10	1648.40	11.679	3.0067	2.1260	0.01719	4.332	3.2833
D	185.00	2819.60	13.472	3.8967	2.2502	0.01693	7.220	15.3664

(i) State the assumptions required for the analysis of variance (ANOVA). State, giving reasons, whether the untransformed data appear to satisfy the assumptions.

(6)

(ii) Assuming that one of these transformations must be chosen, which appears to be the best? Explain why you selected it. Is there evidence that a better transformation could be found? Explain your answer.

(5)

(iii) Produce the ANOVA table for the transformation chosen. The (corrected) total sums of squares for the various transformations are shown below.

	Transformation			
	None	$\sqrt{y}$	$\log_{10}(y)$	$\exp(y/100)$
Corrected total SS	111329	201.189	1.18743	322.795

For the selected transformation, does the analysis suggest the herbicides have different effects? Explain your answer.

(6)

(iv) Alternative methods of checking the validity of an analysis of variance model make use of residuals. Explain how a residual is calculated, and briefly describe one way in which a plot of residuals can be used to help check the validity of a model.

(3)