EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY
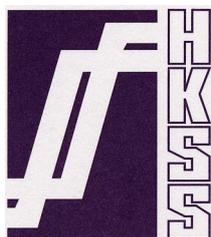
HIGHER CERTIFICATE IN STATISTICS, 2006

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 $\begin{pmatrix} n \\ r \end{pmatrix}$ *is the same as* ^nC_r.

1   **HC Paper III 2006**

1. An experiment was carried out on random samples of steel to assess whether increasing the tempering temperature from 200°C to 250°C increases the failure stress. The samples were treated and tested and gave the following failure stresses (measured in units of 10 MegaPascals).

	Tempering at 200°C	*Tempering at 250°C*
	66	54
	49	63
	58	43
	77	56
	39	47
	51	97
	46	85
	91	
Sample mean	59.6	63.6
Sample standard deviation	17.4	20.1

(i) Carry out a t test, stating carefully your null and alternative hypotheses.

(6)

(ii) State the assumptions you have made in carrying out the t test and provide any suitable graphical evidence relevant to these assumptions.

(4)

(iii) Re-analyse the data using a suitable non-parametric test, once again stating carefully your null and alternative hypotheses.

(7)

(iv) Summarise your conclusions and comment on the methods used in your analysis.

(3)

2. The tensile strength of a synthetic fibre used to make cloth for men's shirts is of interest to a manufacturer. It is suspected that the strength is affected by the percentage of cotton in the fibre. Random samples of material with cotton percentage 15, 20, 25, 30 and 35 were taken and five pieces from each material were strength tested. The recorded strengths are shown below.

15%	20%	25%	30%	35%
7	14	15	12	7
7	18	18	17	10
15	17	20	12	11
11	16	18	18	15
9	18	19	18	11

(i) Complete the analysis of variance table below and state your conclusion about the effect of the percentage of cotton on the strength of the material.

```
One-way ANOVA:  15%, 20%, 25%, 30%, 35%

Source  DF       SS       MS      F
Factor   *    262.64   65.66     *
Error    *       *        *
Total    *    404.64
```

(6)

(ii) A standardised residual can be defined as a residual divided by its standard error. Using this definition, the standardised residuals are as shown in the table below. Construct a plot of the standardised residuals versus fitted values, and comment on the assumptions underlying the analysis of variance.

15%	20%	25%	30%	35%
−1.17	−1.09	−1.26	−1.43	−1.59
−1.17	0.59	0.00	0.67	−0.34
2.18	0.17	0.84	−1.43	0.08
0.50	−0.25	0.00	1.09	1.76
−0.34	0.59	0.42	1.09	0.08

(7)

(iii) Use the estimate of the error variance to calculate 95% confidence intervals for the mean strength at percentage levels 20, 25 and 30. Present your results graphically. What would you recommend as the percentage of cotton required to produce maximum strength?

(5)

(iv) If there were resources to carry out a few more strength tests, describe briefly what you would do.

(2)

3

3. (i) Explain what is meant by a *simple random sample*.

(1)

(ii) Describe the conditions under which *systematic sampling* might be used.

(3)

(iii) A survey is carried out in order to investigate the frequency with which students of a university use the library. Regular use is defined as at least two visits per week during term time.

It is decided to question the first 200 students passing through the main entrance of the university after a specified time. The students are asked whether they use the library regularly and also to state which faculty they belong to. The results are shown below, further classified according to the sex of the respondent.

Faculty	Regular users		Non-Regular users	
	male	*female*	*male*	*female*
Engineering	18	7	20	5
Business	29	13	20	8
Arts	8	13	8	6
Informatics	10	2	22	11

(a) What would be your estimate of p, the proportion of students who regularly use the library?

(1)

(b) Ignoring the sex of respondents, test whether there is any association between faculty and library use.

(7)

(c) Ignoring faculties, calculate an approximate 95% confidence interval for the difference between the overall proportions of male and female students who are regular users of the library. Is there evidence of a difference between the proportions of males and females who regularly use the library?

(5)

(d) Comment on the design of this survey and any factors not taken into account in the selection of students to survey.

(3)

4

4. The following data refer to the sheep population in millions, y, of a certain country over 20 consecutive years, x, together with some of the values for *MA*, the 5-point moving average of y.

x	y	*MA*	x	y	*MA*
1	20	–	11	13	14.2
2	19	–	12	14	13.8
3	18	18.2	13	13	
4	17	17.8	14	14	
5	17	17.6	15	15	
6	18	17.4	16	16	
7	18	17.2	17	17	
8	17	16.8	18	17	
9	16	15.8	19	16	–
10	15	15.0	20	16	–

(i) Estimate the trend in the data using

    (a)   *MA* (complete as appropriate the values for *MA* in the table),

    (b)   linear regression (summary statistics are given below).

$$\sum x_i = 210 \quad \sum y_i = 326 \quad \sum x_i^2 = 2870 \quad \sum x_i y_i = 3305$$

(6)

(ii) Plot a graph of the data and the regression and *MA* estimated trends. Comment on how well each fits the data. Comment on the use of an unweighted, as opposed to a weighted, moving average.

(10)

(iii) State the assumptions made about the error terms in simple linear regression. Comment on how well the data support these assumptions.

(4)

5

5.   The survival time of patients treated for a certain critical illness is a random variable, T, which is assumed to follow a distribution having probability density function

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

(i)   Show that the survivor function (the probability of surviving beyond time t) is

$$S(t) = (1 + \lambda t)e^{-\lambda t}.$$
(3)

(ii)   For a random sample of n observations, $t_1, t_2, \ldots, t_n$, drawn from the defined distribution, derive the maximum likelihood estimator of λ.

Survival times (in months) of a group of ten patients were

16, 23, 35, 44, 51, 55, 67, 97, 127, 192.

Calculate $\hat{\lambda}$, the maximum likelihood estimate of λ, for the above data.
(6)

(iii)   Using $\hat{\lambda}$, estimate the probability that a patient treated for the illness survives longer than 240 months.
(3)

(iv)   It is thought that a function which provides a good estimate of the cumulative distribution function (cdf) for a random sample of data is $F^*(x)$, defined as $F^*(x) = (i - 0.5)/n$, where i is the number of sample values less than or equal to x. Thus, for the data above, $F^*(16) = 0.5/n$, $F^*(23) = 1.5/n$, and so on.

Plot $F^*(x)$ against x, for the given survival times, on a graph. Using $\hat{\lambda}$ as an estimate of λ, calculate the estimated cumulative distribution function values for the given survival times, and plot them on the same graph.

Use your graph to comment on how well this set of data is represented by the fitted model. Does this have any consequences for your answer to part (iii)?
(8)

6

Turn over

6.   In an experiment in which the subject of interest was the effect on weight gain of different amounts and types of protein, six groups of 10 male rats each were given diets that contained two levels of protein and three different types of protein. The mean weight gains (in grams) are shown in the table below.

		Protein type		
		A	*B*	*C*
Protein level	*High*	100.0	89.9	99.5
	Low	79.2	83.9	78.7

(i)   Explain what is meant by *interaction*. Draw a suitable plot for the data above and comment on whether there appears to be any interaction between the type and level of protein used in the animals' diet.

(5)

(ii)   Based on the animals' individual weights, the analysis of variance for this experiment is shown below. Complete the table. Explain carefully the interpretation of the *F* ratios. Comment on the relationship between your conclusions and the plot in (i).

```
Analysis of Variance

Source            DF          SS     MS      F
Level             *       3776.3      *      *
Type              *         82.5      *      *
Level*Type        *        730.1      *      *
Error             *            *      *
Total             *      16174.9
```

(12)

(iii)   Calculate the proportion of overall variation in the data explained by the fitted model, and the estimated underlying residual variance. Comment on your answers.

(3)

Turn over

7. For a study into the density of population around a large city, a random sample of 10 residential areas was selected, and for each area the distance from the city centre and the population density in hundreds per square kilometre were recorded. The following table shows the data and also the log of each measurement.

distance, x (km)	population density, y	log x	log y
0.4	149	−0.916	5.004
1.0	141	0.000	4.949
3.1	102	1.131	4.625
4.5	46	1.504	3.829
4.7	72	1.548	4.277
6.5	40	1.872	3.689
7.3	23	1.988	3.135
8.2	15	2.104	2.708
9.7	7	2.272	1.946
11.7	5	2.460	1.609

(i)   By plotting three separate graphs, decide which of the following regressions is best represented by a straight line.

(a) y on x (b) y on $\log x$ (c) $\log y$ on x

(7)

(ii)   On the basis of the regression results **on the next page**, which regression do you consider to be best? Justify your answer by reference to the diagnostic criteria given in the output and relating these to your plots in (i). Would you consider regressing $\log y$ on $\log x$? If not, why not?

(5)

(iii)   For the model you consider to be best in (ii), obtain an expression for y in terms of x.

(3)

(iv)   Using your chosen model, estimate the density of the population at a distance of 5 km from the city centre.

(2)

(v)   State any reservations you have about using the model to predict population density.

(3)

The regression results for this question are on the next page

Turn over

Regression results for question 7

Regression Analysis: y versus x

```
The regression equation is   y = 140 - 14.0x

Predictor      Coef   SE Coef      T      P
Constant     139.70     11.12  12.56  0.000
x            -13.958     1.663  -8.39  0.000

S = 18.2834   R-Sq = 89.8%   R-Sq(adj) = 88.5%

Observation 10 has an unusually large positive residual
```

Regression Analysis: y versus logx

```
The regression equation is y = 127 - 48.0logx

Predictor      Coef   SE Coef      T      P
Constant     126.990    9.147  13.88  0.000
logx         -47.980    5.293  -9.07  0.000

S = 17.0492   R-Sq = 91.1%   R-Sq(adj) = 90.0%

Observation 1 has an unusually large negative residual
```

Regression Analysis: logy versus x

```
The regression equation is logy = 5.41 - 0.322x

Predictor       Coef   SE Coef        T      P
Constant      5.4133    0.1621    33.40  0.000
x            -0.32157  0.02425   -13.26  0.000

S = 0.266544   R-Sq = 95.6%   R-Sq(adj) = 95.1%
```

Turn over

8. In a project designed to see whether testing procedures in different laboratories give similar results, fatigue tests at three different strain levels (level 1 lowest, level 3 highest) were carried out on samples from the same batch of a composite material. The tests were carried out at 9 different laboratories across Europe. The results (cycles to fatigue) from each laboratory and the means of the samples at each strain level are shown in the table below.

laboratory	strain level 1 results	mean (1)	strain level 2 results	mean (2)	strain level 3 results	mean (3)
A	7335, 6882, 8353	7523	1336, 1693	1515	690, 735	712.5
B	3512, 4000, 4300	3937	977, 1152	1065	428, 460, 473	453.7
C	3822, 5558, 6910	5430	1516, 1607, 1650	1591	630, 675	652.5
D			1740, 1852	1796	447, 718, 935	700
E	4382, 6239, 8930	6517	1488, 1501, 1516	1502	674, 681, 781	712
F	4030, 4138, 4202	4123	1095, 1205, 1290	1197	465, 380, 395	413.3
G	5000, 5245, 5452	5232	1031, 1156, 1238	1142	380, 408, 454	414
H	2510, 2604, 2811, 2900, 2986	2762	738, 771, 864, 883	814	303, 325, 329	319
J	2247, 3800, 5267	3771	957, 1156, 1202	1105	225, 487	356

Two questions are of interest:

(1) Are there substantial differences between the measurements obtained at different laboratories?

(2) Can a model be found to predict the cycles to fatigue at strain levels other than those tested?

Write a preliminary report on your general conclusions about the questions posed in (1) and (2), based on the tabulated data, and supported by suitable graphical evidence. (You are advised to avoid spending excessive time on detailed statistical analysis or repetitive calculations.)

Express your report in a way that makes it accessible to non-technical readers.

(20)