EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2014

MODULE 5 : Topics in applied statistics

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 12 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. Parkinson's Disease (PD) is a common neurological disorder, caused by the loss of brain cells that produce the chemical dopamine. A study was carried out to compare brain activity in PD sufferers and controls. A representative sample of 42 PD patients, at various stages of the disease, underwent brain imaging using Single Photon Emission Computerised Tomography. 14 controls of similar age to the patients, but with no known neurological disorder, were also recruited into the study and imaged in the same way.

A measure of brain activity was obtained for each of the Striatum (Sr), Caudate (Ca) and Putamen (Pu) regions on the left (L) side and right (R) side of the brain. These six regions are known to be important in dopamine production. This measure of activity was dimensionless but larger values indicated more brain activity.

The Hoehn and Jahr (HY) disease stage was also recorded for each PD sufferer. This is a clinical indicator of the progress of the disease, which proceeds from Stage 1 to Stage 5 as the condition gets worse. In this study, Stage 0 was used to indicate normal controls.

The computer output displayed **on the next page** is from an analysis of the data for all 56 persons in the study.

(i) What does the sample correlation matrix (Table 1) reveal about the brain activity levels recorded in this study?

(3)

(ii) A principal component analysis of the data was carried out, based on the sample correlation matrix. Discuss some possible objectives of this analysis, and its limitations.

(5)

(iii) A scree plot is given in Figure 1. What can be concluded from this plot? Having seen this plot, how would you proceed with the analysis?

(5)

(iv) Table 2 lists the loadings of the first two principal components. Use these loadings to interpret each of the two components.

(4)

(v) Figures 2 and 3 show box and whisker plots (boxplots) of the first two principal component scores for the subjects in the study, according to their HY stage. Comment on these plots.

(3)

Output for Question 1 is on the next page

Table 1: Sample correlations

	SrL	SrR	CaL	CaR	PuL	PuR
SrL	1.00	0.93	0.98	0.92	0.99	0.91
SrR	0.93	1.00	0.91	0.98	0.92	0.99
CaL	0.98	0.91	1.00	0.93	0.95	0.88
CaR	0.92	0.98	0.93	1.00	0.89	0.95
PuL	0.99	0.92	0.95	0.89	1.00	0.91
PuR	0.91	0.99	0.88	0.95	0.91	1.00

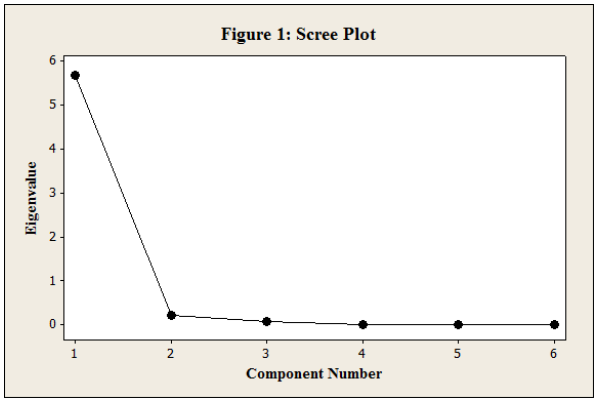
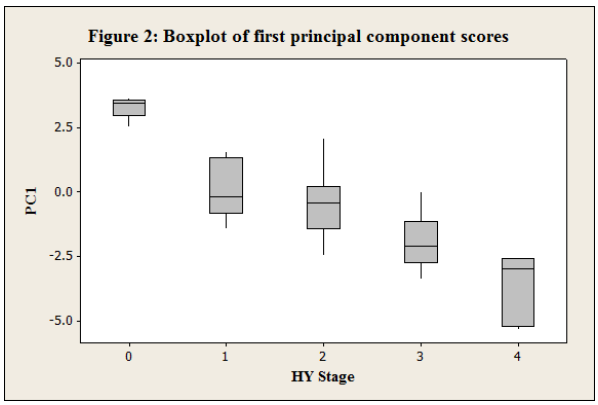
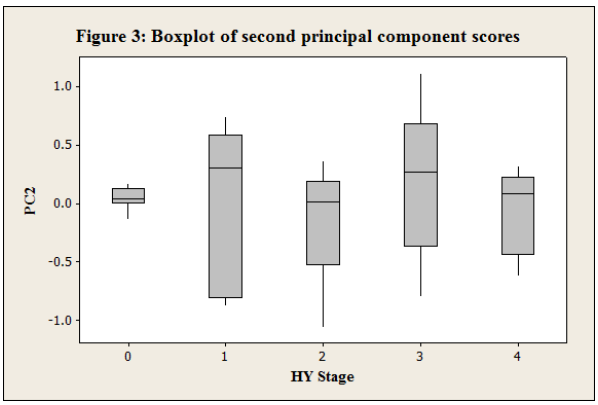


Table 2: Loadings for the first two principal components

	SrL	SrR	CaL	CaR	PuL	PuR
PC1	0.412	0.412	0.406	0.407	0.407	0.406
PC2	-0.406	0.413	-0.404	0.361	-0.411	0.449





2. (i) Briefly outline the purpose of *discriminant analysis* and give reasons for using it.

(3)

(ii) Briefly outline the purpose of *cluster analysis* and give reasons for using it. Discuss ways in which cluster analysis and discriminant analysis are similar and ways in which they are different.

(6)

(iii) In the context of discriminant analysis, explain why estimates of misclassification probabilities might be biased if a discriminant function were estimated from and tested on the same data. Discuss possible approaches for reducing or eliminating this bias.

(8)

(iv) Consider the problem of discriminating between two groups using just one continuous explanatory variable. Explain why a significance test might show that the variable is significantly different on average in the two groups but discriminant analysis suggests that the variable is not useful for classifying objects.

(3)

3. (i) In a small clinical trial, 21 leukaemia patients were treated with a new drug and their time in remission (in weeks) following treatment was recorded. The results are given below, where * denotes that the patient was still in remission when the study finished.

6*	6	6	6	7	9*	10*	10	11*	13	16
17*	19*	20*	22	23	25*	32*	32*	34*	35*	

Use the Kaplan-Meier method with these data to estimate the survivor function, $S(t)$, for the length of remission, t, in leukaemia patients treated with this drug.

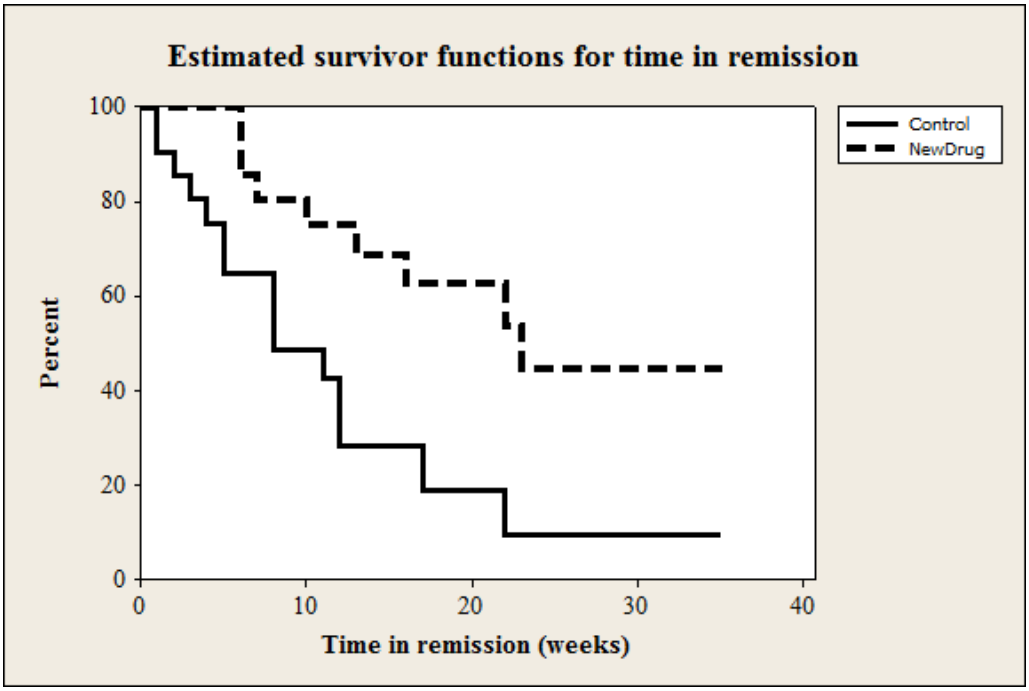
(8)

(ii) Using Greenwood's formula, calculate the standard error for the Kaplan-Meier estimate of $S(t)$ at 10 weeks. Hence provide an approximate 95% confidence interval for $S(t)$ at 10 weeks. Interpret and comment on this interval.

(9)

(iii) In the study, the patients treated with the new drug were compared with a control group of 21 similar patients. The estimated survivor functions for the two treatment conditions are shown in the figure below. The log rank test was carried out to compare the survivor functions for the two treatments, with the result $p = 0.006$. Interpret this result, together with the plotted survivor functions.

(3)



4. Data were collected in order to investigate the length of time that a cohort of 238 heroin addicts (the subjects) could be maintained on methadone treatment. Methadone is a substitute for heroin and is therefore only helpful to addicts for as long as they are taking it. In the study, time maintained on methadone (days) was used as a right-censored response variable. All the subjects were assessed at the same central unit and then referred to one of two clinics (Clinics 1 and 2) for maintenance. The maximum daily dose of methadone (mg) given to each subject was recorded. Among several subject characteristics, whether or not the individual had a prison record was also noted.

(i) It was decided to fit a Cox proportional hazards model to the data, with maximum dose of methadone, clinic and prison record as explanatory variables. Write down the form of this model, interpreting clearly each of the terms in it.

(5)

(ii) The model was fitted and the results shown below were obtained. What can be concluded about the effects of the three explanatory variables?

(6)

	Coefficient	*Standard Error*
Dose (mg)	–0.354	0.006 38
Clinic: Clinic 2	–1.01	0.215
Prison record: Yes	0.327	0.167

(iii) Obtain a 95% confidence interval for the hazard ratio of a subject with a prison record relative to a subject without a prison record, assuming the two subjects are maintained in the same clinic with the same maximum dose of methadone.

(3)

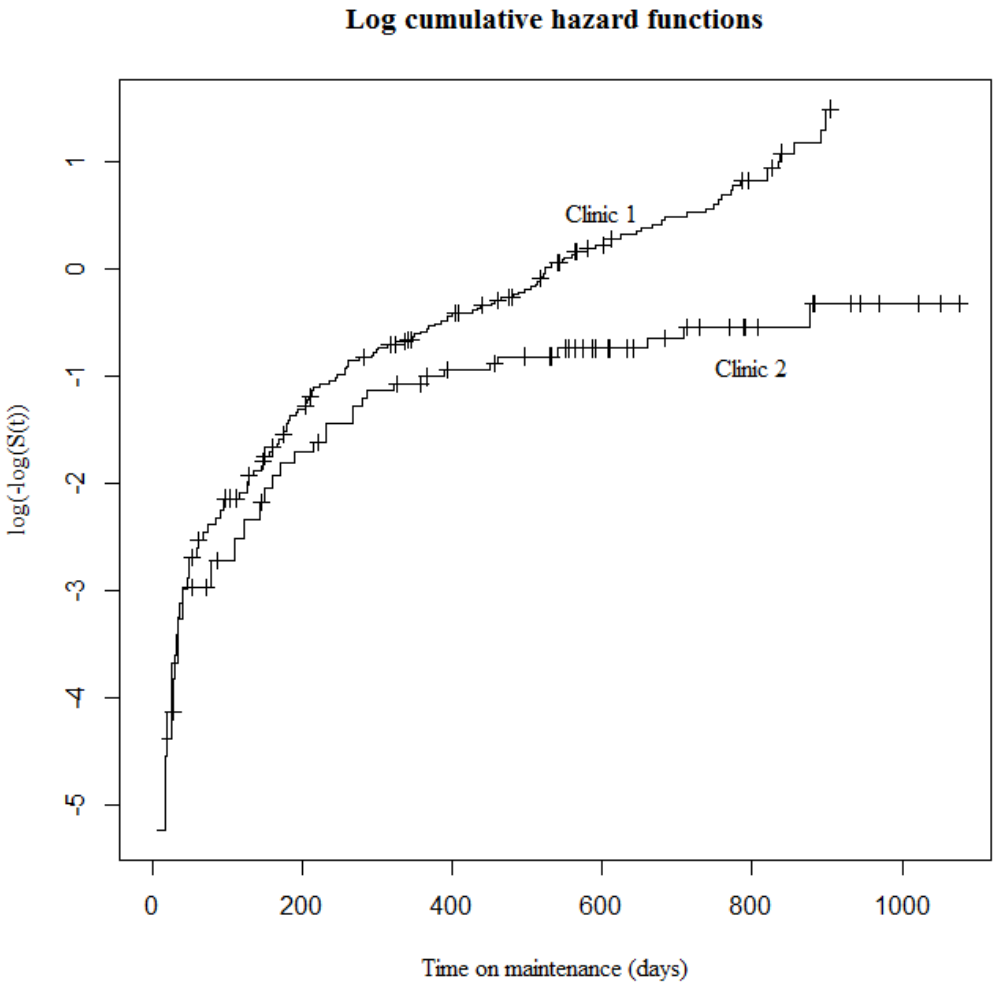
(iv) Obtain a 95% confidence interval for the hazard ratio of a subject with a maximum dose of 80 mg of methadone relative to a subject with a maximum dose of 60 mg, assuming the two subjects are maintained in the same clinic and have the same prison record.

(4)

(v) The subjects were divided into two groups according to the clinic where they were maintained. The estimated log cumulative hazard functions for the two groups are compared in the figure **on the next page**. What do you conclude from this plot?

(2)

Figure for part (v) is on the next page


Log cumulative hazard functions
log(-log(S(t)))
1
0
-1
-2
-3
-4
-5
Clinic 1
Clinic 2
0
200
400
600
800
1000
Time on maintenance (days)

5. A randomised clinical trial was carried out to compare the drug Metoprolol against a placebo for the treatment of patients who had suffered a heart attack. The outcome of interest was death within one year of commencing treatment. The outcomes are shown in the following table.

No. of Deaths/Total No. of Patients (%)

	Placebo	*Metoprolol*
Aged 40–64 years	26/453 (5.7%)	21/464 (4.5%)
Aged 65–69 years	25/174 (14.4%)	11/165 (6.7%)
Aged 70–74 years	11/ 70 (15.7%)	8/ 69 (11.6%)

(i) For patients aged 40–64 years, obtain an estimate of the odds ratio for dying within one year after starting treatment on Metoprolol relative to placebo. Obtain an approximate 95% confidence interval (CI) for this odds ratio, and comment on your answer.

(8)

(ii) The equivalent results for the other age groups are:

ages 65–69 years, estimated odds ratio = 0.43 (95% CI, 0.20 – 0.90);

ages 70–74 years, estimated odds ratio = 0.70 (95% CI, 0.26 – 1.87).

What do you conclude about the effectiveness of Metoprolol for different age groups?

(4)

(iii) Use the Mantel-Haenszel procedure to obtain a pooled estimate of the odds ratio, adjusted for age. Interpret this statistic.

(6)

(iv) Give a reason for preferring the Mantel-Haenszel approach to simply obtaining an estimate of the odds ratio from the total counts across the three age groups.

(2)

6. (i) Explain clearly the purpose of *standardisation* with respect to death rates, and distinguish between *direct* and *indirect* standardisation.

(5)

The data below summarise the age distributions, and the associated prevalence of coronary heart disease (CHD) mortality, for males and females in the United Kingdom in 2010. The table also gives the hypothetical European Standard Population (ESP), which is the same for both sexes and is used for comparisons between countries and over time, and the age distribution for Scotland, a subpopulation of the United Kingdom.

	UK population		**CHD Deaths**		**ESP**	**Scottish population**	
Age	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*
< 35	13 918	13 316	102	36	50	1 119	1 079
35–44	4 378	4 456	681	166	14	348	377
45–54	4 215	4 333	2 539	586	14	370	398
55–64	3 599	3 743	5 899	1 495	11	318	334
65–74	2 572	2 827	9 952	4 084	7	220	254
75+	1 961	2 944	27 418	27 610	4	155	250
Total	30 643	31 619	46 591	33 977	100	2 530	2 692

All population figures are counts in thousands; CHD deaths are counts.

(ii) Calculate the crude rates for the prevalence of CHD mortality (in deaths per 100 000) separately for males and females in the United Kingdom in 2010.

(2)

(iii) Calculate the direct adjusted rate for the prevalence of CHD mortality among males in the United Kingdom in 2010 using the European Standard Population as the standard population.

(5)

(iv) In 2010, CHD accounted for 8138 deaths in Scotland (4599 males and 3539 females). Calculate the standardised mortality ratio for males in Scotland in 2010, using the male age distribution for the United Kingdom in 2010 as the standard population.

(5)

(v) You may assume that the direct adjusted rate for prevalence of CHD mortality among females in the United Kingdom in 2010 is 54.6 deaths per 100 000, and the standardised mortality ratio for the prevalence of CHD mortality for females in Scotland in 2010 is 1.21. Comment on the conclusions that may be drawn from comparisons between these values and the results obtained in parts (ii) – (iv).

(3)

7. (i) In a population of N units, the binary variable X takes the value 1 in $N\theta$ units and the value 0 in the other $N(1-\theta)$ units (where $0 < \theta < 1$). Without loss of generality, $N\theta$ and $N(1-\theta)$ can both be assumed to be integers. A simple random sample of n units is taken without replacement, and θ is estimated by $p = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_i is the value of X for the ith sampled unit. Show that p is an unbiased estimator of θ with approximate variance $\frac{\theta(1-\theta)(1-f)}{n}$, where f is the sampling fraction.

(6)

[You may use without proof the result that the random variable Y with probability distribution $P(Y=y) = \binom{K}{y}\binom{M-K}{m-y}\Big/\binom{M}{m}$, for $y = 0, \ldots, K$, has expected value $m\frac{K}{M}$ and variance $m\frac{K}{M}\left(1-\frac{K}{M}\right)\frac{M-m}{M-1}$.]

(ii) A population consists of 20 000 large companies and 80 000 small companies, all of which maintain a presence on the worldwide web. A pilot study suggests that approximately 90% of large companies have adequate internet security but only approximately 50% of small companies do. A sample-based study is to be conducted with the aim of estimating the overall proportion of companies that have adequate internet security. The sample will be stratified by size of company (small or large) and the total sample size will be 1000. It may be assumed that the cost of sampling a company is the same in the two strata.

(a) Write down formulae for the stratified sampling estimator of the overall proportion and its estimated variance. You should ignore the finite population correction.

(2)

(b) Suppose now that the percentage estimates from the pilot study are close to the true values. Calculate the required number of companies in the sample for each of the two strata, using:

(1) proportional allocation;

(2) optimal allocation.

(5)

(c) Calculate and compare the relative efficiencies of these two methods of allocation with that of a simple random sample of 1000 companies. (Continue to ignore the finite population correction.) Which sampling method would you recommend for use in the proposed study, and why?

(7)

8. (a) (i) Explain briefly how a sample is obtained in one-stage cluster sampling and in two-stage cluster sampling.

(7)

(ii) Describe the main similarities and differences between *clusters* appropriate for use in cluster sampling and *strata* appropriate for use in stratified random sampling.

(5)

(b) A widely-scattered, rural population in a developing country is divided into 200 area segments, each containing 5 households. A simple random sample of 20 segments is selected from the population, all households in each segment are visited and the number of people in each of these households is recorded. Let y_{ij} be the number of people in the jth household of the ith sampled segment. The results are summarised as follows.

$$\sum_{i=1}^{20}\sum_{j=1}^{5} y_{ij} = 556 \qquad \sum_{i=1}^{20}\left(\sum_{j=1}^{5} y_{ij}\right)^2 = 16\,062$$

Give one reason why a cluster design was chosen in preference to simple random sampling for this study. Obtain an approximate 95% confidence interval for the total number of people in the population.

(8)

[You may assume without proof that, if $\hat{T}$ is an appropriate estimator of the total population size, then its estimated variance is

$$N^2\left(1-\frac{n}{N}\right)\frac{1}{n(n-1)}\left\{\sum_{i=1}^{n} t_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\right)^2\right\}$$

where there are N clusters in the population, n clusters in the sample, and t_i people in the ith sampled cluster.]

BLANK PAGE