EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2011

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 13 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) For what purposes would you use a linear discriminant function in a two-group problem? Discuss the conditions required for this method of analysis to be appropriate.

(4)

(ii) In a study of psychology students, aimed at predicting their academic success, two variables were considered: x_1, the GRE-Q score (a measure of quantitative ability), and x_2, the number of hours of psychology courses taken previously. Random samples of 30 successful students and 20 unsuccessful students were taken. The results were summarised as follows. For the successful group, the mean vector, $\bar{\mathbf{x}}_S$, and sample covariance matrix, $\hat{\mathbf{\Sigma}}_S$, were as follows.

$$\bar{\mathbf{x}}_S = \begin{bmatrix} 548.5 \\ 28.7 \end{bmatrix} \qquad \hat{\mathbf{\Sigma}}_S = \begin{bmatrix} 8955 & -13 \\ -13 & 127 \end{bmatrix}$$

For the unsuccessful group, the mean vector, $\bar{\mathbf{x}}_U$, and sample covariance matrix, $\hat{\mathbf{\Sigma}}_U$, were as follows.

$$\bar{\mathbf{x}}_U = \begin{bmatrix} 498.8 \\ 14.4 \end{bmatrix} \qquad \hat{\mathbf{\Sigma}}_U = \begin{bmatrix} 9323 & 282 \\ 282 & 72 \end{bmatrix}$$

(a) Use this information to provide a linear discriminant function to distinguish between these two groups. You may assume that the two types of possible error are considered to be equally important.

(7)

(b) Stating any assumptions you make, estimate the probability of classifying a future psychology student to the wrong group.

(5)

(c) Consider two new psychology students, A and B, where A has $x_1 = 525$ and $x_2 = 22$, while B has $x_1 = 530$ and $x_2 = 23$. Use your results to predict whether A will be successful or unsuccessful. If you were now asked to predict the outcome for B, would you be more or less confident about your prediction? Explain your answer.

(4)

2. (i) What is the motivation behind Principal Components Analysis (PCA)? What are the drawbacks of PCA?

(5)

(ii) Data are available on the crime rates per 100,000 people in each of 72 cities in the United States of America in 1994. The variables represent rates for the following crimes.

- Murder
- Rape
- Robbery
- Assault
- Burglary
- Larceny
- Motor Vehicle Theft (MVT)

The output **on the next three pages** gives some results from a PCA based on the sample correlation matrix for these data. Making reference to this output, answer the following questions.

(a) What does the scatterplot matrix (Figure 1) reveal about the relationships between the seven variables?

(3)

(b) How many components would you choose to retain from this PCA? Referring to specific parts of the output, give reasons for your choice.

(3)

(c) Table 2 provides the principal component coefficients. Figure 2 shows a scatterplot of the scores for 5 of the 72 cities on the first (x-axis) and second (y-axis) principal components. Give interpretations of the first two principal components. Discuss the levels of crime in Santa-Ana, Corpus-Christi and Newark.

(5)

(d) Name one other multivariate data analysis technique that you might use to explore these data, and give reasons for your choice.

(4)

Output for question 2 is on the next three pages

Output for question 2 (page 1 of 3)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Std dev.	1.948	1.095	0.877	0.714	0.577	0.461	0.425
Var. prop.	0.540	0.170	0.110	0.073	0.048	0.031	0.028
Cum. prop.	0.540	0.710	0.820	0.893	0.941	0.972	1.000

Table 1. Principal components information

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Murder	0.370	–0.339	0.202	–0.717	–0.277	0.220	–0.262
Rape	0.249	0.466	0.782	0.159	0.010	0.267	0.021
Robbery	0.426	–0.387	0.079	0.012	0.194	–0.147	0.776
Assault	0.434	0.042	–0.282	0.021	0.767	0.118	–0.358
Burglary	0.449	0.238	0.015	–0.032	–0.242	–0.794	–0.220
Larceny	0.276	0.605	–0.492	–0.209	–0.264	0.303	0.331
MVT	0.390	–0.302	–0.134	0.644	–0.399	0.351	–0.204

Table 2. Principal component coefficients

Turn over

Output for question 2 (page 2 of 3)

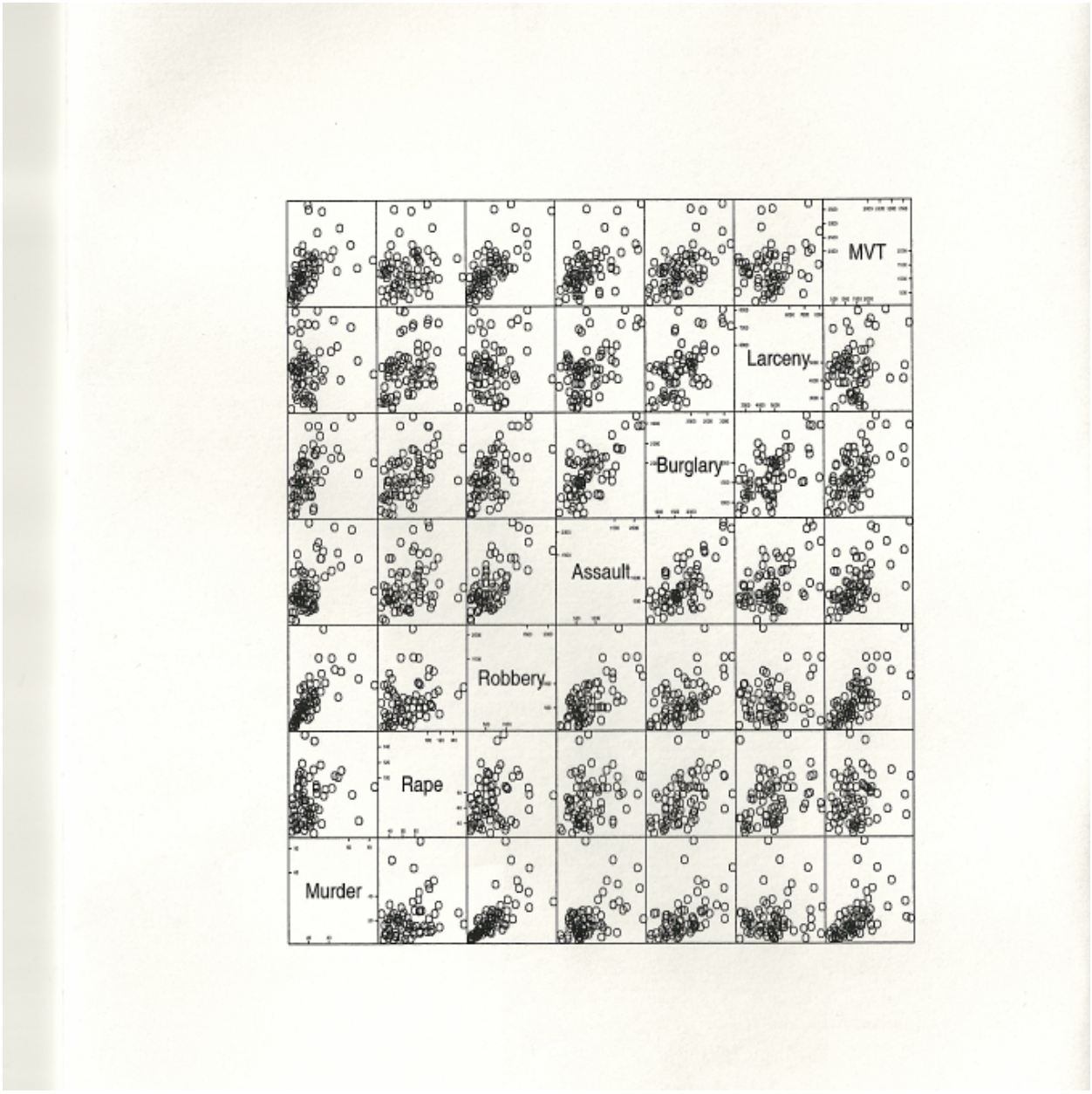


Figure 1. Scatterplot of US crime data

Turn over

Output for question 2 (page 3 of 3)

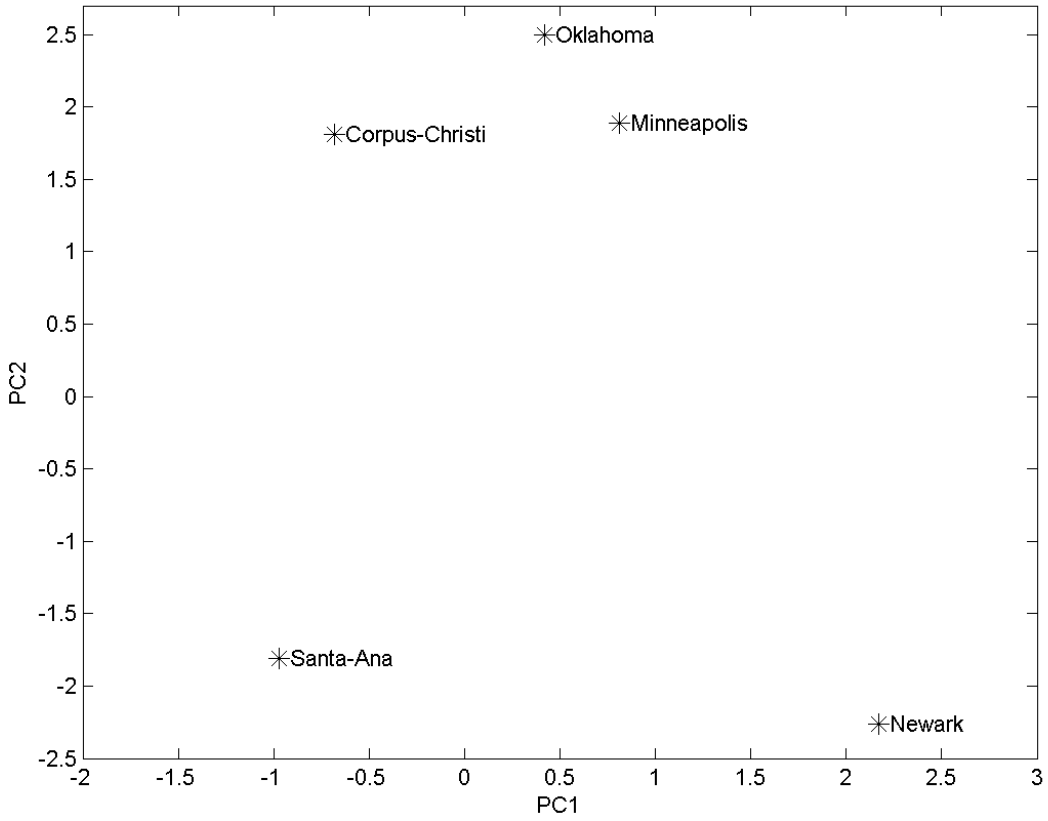


Figure 2. Graph of Principal Component 2 (PC2) against Principal Component 1 (PC1)

Turn over

3. (i) Define the *survivor function* and the *hazard function* for a random variable T measuring lifetime. Explain how the two functions are related.

(3)

(ii) What is a *right-censored* observation? Give two different examples of ways in which a right-censored observation might arise.

(3)

(iii) If a right-censored observation is denoted by *, derive the Kaplan-Meier estimate of the survivor function for the following data.

1* 3 4* 5 5 6* 7* 7 7 8

Plot the result on an appropriate graph.

(5)

(iv) The survival times (months) of fifteen patients aged 21 – 40 suffering from a skin tumour called melanoma, who were all treated with either the BCG or the *c. parvum* vaccine, are shown below.

BCG	8	17*	17*	19	24*	34*			
c. parvum	7	8	8*	12*	16*	18*	21*	24*	27*

Using the logrank test, test the hypothesis that there is no difference between the distributions of the survival times when patients are treated with BCG and when they are treated with *c. parvum*.

(9)

4 (i) Write down the proportional hazards model with time-independent covariates Z, and interpret each term in your model. Define the *hazard ratio* in the context of this model. What assumptions does this model make about the effect of the covariates on the hazard function?

(6)

(ii) A clinical trial was conducted to study the effect of a new cancer treatment. Eligible patients were randomised into two groups, one treated with the new treatment and the other with the current treatment. Data from the trial were analysed using a proportional hazards model. Two factors were included: TREAT, taking the value 0 (current) or 1 (new); SEX, taking the value 0 (male) and 1 (female). The following computer output shows the results from the model including the main effects of the two factors and the interaction between them.

```
n = 400

                coef      exp(coef)  se(coef)
TREAT           0.644     1.90       0.257
SEX             0.289     1.33       0.207
TREAT:SEX      -0.235     0.79       0.301

Likelihood ratio test = 15.4 on 3 df, p = 0.00154
```

(a) What can you conclude about the differences in effect between the two treatments and the two sexes?

(6)

(b) Use the fitted model to estimate the hazard ratio for a female patient taking the new treatment compared to a male patient taking the current treatment.

(3)

(c) The graphs **on the next page** show the Kaplan-Meier estimates of the survivor function for the Current treatment (solid line) and the New treatment (dashed line) within each sex. Referring to these graphs, discuss whether the proportional hazards model is appropriate for each group of patients. How would this affect your answer to part (b)?

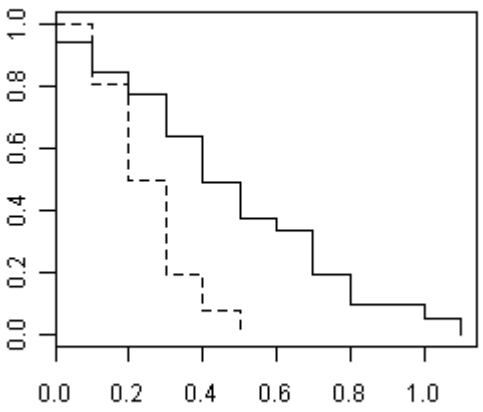
(5)

Graphs for question 4 are on the next page

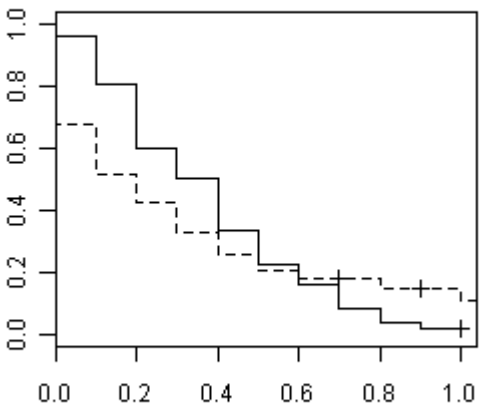
Turn over

Graphs for question 4

Male



Female



Turn over

5. (i) Distinguish between a *cohort study* and a *case-control study* in medical trials, and discuss how controls should be chosen in a case-control study.

(7)

(ii) A company in the health care industry has developed a new version of a diagnostic product used to screen patients for a bacterial infection. Fifty pairs of patients, whose symptoms suggested that they were suffering from the infection, were chosen to be as similar as possible to one another within pairs. One of each pair, chosen at random, was examined using the old version of the product, and the other using the new version. The following table summarises the results of the screening.

		New product	
		Positive	*Negative*
Old product	*Positive*	20	12
	Negative	2	16

(a) Explain briefly why pairing of patients is useful in a study of this type.

(2)

(b) Test at the 5% level of significance the null hypothesis that the true proportion of positives is the same for the two versions of the product. Explain your choice of test.

(6)

(c) Construct an approximate 95% confidence interval for the difference between the true proportions of positives. Describe the approximations made.

(5)

6. Mortality data for coronary heart disease in a recent year, in the United Kingdom as a whole and in Scotland alone, were as follows.

Age	**Scotland**		**United Kingdom**	
	Population (thousands)	*Deaths*	*Population (thousands)*	*Deaths*
Under 35	2513	18	28226	192
35 – 44	701	185	7932	1595
45 – 54	580	751	6593	6035
55 – 64	537	2346	5814	19515
65 – 74	441	4886	5075	46369
75 and over	328	8680	4009	97473
Total	5100	16866	57649	171179

(i) Explain why it is important to standardise death rates when comparing numbers of deaths due to coronary heart disease in different geographical regions. Explain the difference between direct and indirect standardisation.

(5)

(ii) Calculate the crude death rates per thousand in Scotland and in the United Kingdom.

(2)

For the rest of this question, use the data for the United Kingdom as the standard population.

(iii) Calculate the age standardised death rate per thousand for coronary heart disease in Scotland. Compare this with the crude death rate.

(5)

(iv) Calculate the standardised mortality ratio for coronary heart disease in Scotland. What information is provided by the calculation?

(5)

(v) Calculate the indirect standardised death rate per thousand for coronary heart disease in Scotland. Comment.

(3)

7. A sample of n items is to be chosen without replacement from a finite population consisting of N items. A simple random sample is defined as one in which each of the $\binom{N}{n}$ different possible samples has an equal probability of being selected.

(a) Show that, by this method of sampling, each item in the population has an equal probability of being chosen to be in the sample.

(2)

(b) A measurement X of a particular characteristic is to be taken on each member of the sample, giving data $x_1, x_2, ..., x_n$. The distribution of X in the population has mean $\overline{X}$ and variance σ^2.

(i) Show that the sample mean, $\overline{x}$, is an unbiased estimator of $\overline{X}$.

(1)

(ii) Define the variance of $\overline{x}$ as $\text{Var}(\overline{x}) = E\left[\left(\overline{x} - \overline{X}\right)^2\right]$, taken over all possible samples. Also define $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2$, so that $(N-1)S^2 = N\sigma^2$. Prove that $\text{Var}(\overline{x}) = \frac{(1-f)S^2}{n}$, where $f = \frac{n}{N}$.

(7)

(iii) Suppose that the mean of a sample of $n = 15$ items is $\overline{x} = 12$, and an unbiased estimate of the population variance is $s^2 = 16$. Obtain a 95% confidence interval for $\overline{X}$ if the population size N is not known. In each of the cases $N = 500$ and $N = 60$, recalculate this interval using the finite-population correction $(1 - f)$. Comment on the effect this correction has on the intervals, and suggest a working rule for when it should be used.

(3)

(c) (i) A scientist needs to use a very small sample of items to estimate the population proportion P that have a particular characteristic. Verify that, if $a = 2$ out of $n = 12$ of the items show the characteristic, then the 95% confidence interval for P using the Normal approximation method is (–0.053, 0.387). Comment on this interval, and explain why intervals like this can arise.

(3)

(ii) An alternative interval, from P_L to P_U, can be calculated using the equations for $P(a \geq 2)$ and $P(a \leq 2)$ respectively. Write down these equations, and verify that $P_L = 0.021$ and $P_U = 0.484$. Comment on this result.

(4)

8. An agricultural research officer working in a developing country has been put in charge of a survey which aims to find out how much of the area under food crops is being used for growing maize. He has been given aerial maps of the country, and information on which regions of the country contain land suitable for growing crops. Farms and villages in the regions are scattered over large areas. The officer has been told that two methods which have worked satisfactorily for economic and social surveys in the main city are stratified random sampling and cluster sampling. He asks your opinion on whether he should use one of those methods and, if so, which one.

(i) Give an example of an economic or social survey in the city for which either of these two methods could be used, and explain carefully the advantages and drawbacks of each method for that survey.

(8)

(ii) Compare and contrast the strengths and weaknesses of these two methods for use in the crop survey. You should make clear what information will be needed in order to use each method, and bear in mind that there is a limited budget for the survey. Which method would you recommend to the research officer, and why?

(7)

(iii) As an afterthought, the officer remarks that a friend of his in the city said that systematic sampling was so much quicker to carry out than random sampling that he preferred to use that sometimes. Explain to the officer what difficulties might arise in collecting data by systematic sampling and in analysing and reporting on the data collected in this way.

(5)