EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2015

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* ${}^nC_r$.

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. Conn's Syndrome is a form of hypertension that has two possible causes: an adenoma (Type A patient), which has to be removed by surgery, and bilateral hyperplasia (Type B patient), which is a more diffuse condition and is treated with drugs. It can be hard to tell whether a patient is Type A or Type B. Researchers investigated a group of 31 sufferers of Conn's Syndrome, recording their age (in years) and the concentrations of the following three chemicals in blood plasma (in meq/l): sodium, potassium, carbon dioxide. All these patients then underwent surgery, which revealed that 20 of them were Type A and the other 11 Type B.

The computer output **displayed on the next page** is from an analysis of the data for all 31 patients in the study.

(i) What does the matrix plot suggest about the potential usefulness of each of these variables for classifying patients as Type A or Type B?

(4)

(ii) A linear discriminant analysis of the data was carried out, with the results shown in Table 1. Explain what is meant by *leaving-one-out cross-validation*. Give one reason for carrying out this procedure as part of discriminant analysis. What do the results in Table 1 suggest about the usefulness of the linear discriminant?

(6)

(iii) The linear discriminant function is

$$y = -20.0 - (0.1 \times \text{age}) + (0.2 \times \text{sodium}) - (3.0 \times \text{potassium}) + (0.6 \times \text{carbon dioxide}),$$

where positive values of y indicate Type A patients. Use it to classify another Conn's Syndrome sufferer, who is 40 years old and has the following test results: sodium 144.1, potassium 3.4, carbon dioxide 25.2 meq/l.

(2)

(iv) List the main statistical assumptions required to justify the use of linear discrimination. In what way may these assumptions be relaxed if quadratic discrimination is used instead of linear discrimination?

(4)

(v) Table 2 gives the results of quadratic discrimination, in a form comparable to Table 1. Compare the results from the two discriminants. Which would you recommend using in practice, and why?

(4)

Output for Question 1 is on the next page

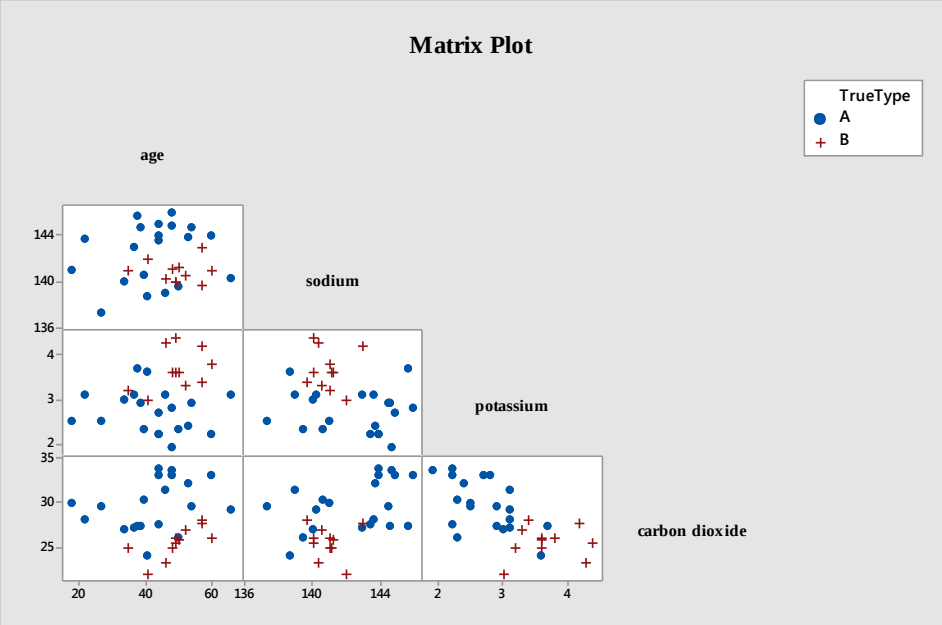

Table 1: Results of Linear Discriminant Analysis

(a) Without Cross-Validation

	True Group	
Put into Group:	A	B
A	17	0
B	3	11
Total	20	11
No. correct	17	11
Proportion	0.850	1.000

Overall Proportion Correct = 0.903

(b) With Cross-Validation

	True Group	
Put into Group:	A	B
A	16	1
B	4	10
Total	20	11
No. correct	16	10
Proportion	0.800	0.909

Overall Proportion Correct = 0.839

Table 2: Results of Quadratic Discriminant Analysis

(a) Without Cross-Validation

	True Group	
Put into Group:	A	B
A	19	0
B	1	11
Total	20	11
No. correct	19	11
Proportion	0.950	1.000

Overall Proportion Correct = 0.968

(b) With Cross-Validation

	True Group	
Put into Group:	A	B
A	17	3
B	3	8
Total	20	11
No. correct	17	8
Proportion	0.850	0.727

Overall Proportion Correct = 0.806

2. (i) Briefly describe the purposes of *cluster analysis*.

(3)

(ii) Suppose that objects are to be clustered on the basis of an observation vector that consists of p continuous measurements. Define the *Euclidean distance* between objects with observation vectors

$$\boldsymbol{x} = [x_1 \quad x_2 \quad \cdots \quad x_p]^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{y} = [y_1 \quad y_2 \quad \cdots \quad y_p]^{\mathrm{T}}.$$

Also define any two of the following alternative similarity measures: *squared Euclidean distance*; *Manhattan* (or *city-block*) *distance*; *Chebyshev distance*; *Mahalanobis distance*. For each of your chosen cases, describe how the overall features of the complete set of distances obtained using that measure would differ from the corresponding set of Euclidean distances.

(5)

(iii) Describe the *single-linkage* and *complete-linkage* approaches to hierarchical clustering. How might the outcome of a cluster analysis differ depending on which of these approaches was used?

(6)

(iv) Describe *k-means clustering*. Mention one advantage and one disadvantage of this method compared with hierarchical procedures.

(6)

3. (i) A random variable, T, has the Weibull distribution with probability density function

$$f(t) = \begin{cases} \lambda\gamma t^{\gamma-1}\exp\{-\lambda t^{\gamma}\}, & t > 0 \\ 0, & t \le 0 \end{cases}$$

for parameters $\lambda > 0$ and $\gamma > 0$.

(a) Derive the survivor function, $S(t)$, of T.

(3)

(b) Show that $\log\{-\log(S(t))\}$ is a linear function of $\log(t)$. State the intercept and slope of the line if $\log\{-\log(S(t))\}$ on the vertical axis is plotted against $\log(t)$ on the horizontal axis.

(5)

(ii) 20 refrigerator motors of a particular type were each tested on an accelerated life test, and their times till first failure (hours) were recorded. The results are listed below, where * denotes that the motor was still functioning properly when the test was brought to an end.

2 4^* 5 5 5 6^* 7 7 7^* 8 8 9^* 11 11 12 12^* 12^* 12^* 16 18^*

(a) Use the Kaplan-Meier method with these data to estimate the survivor function, $S(t)$, for the time to first failure of a motor of this type.

(4)

(b) Referring to the result from part (i)(b), use a suitable graphical method to investigate whether or not these data come from a Weibull distribution.

(5)

(c) Draw a straight line through the points on your graph by eye and use it to estimate the parameters, λ and γ, of a Weibull distribution fitted to these data.

(3)

4. 154 subjects with burns were monitored in a study of a new treatment to prevent burn wounds becoming infected. 70 of the subjects were given standard care (Treat = 0) while the other 84 had additional care thought to make infection less likely (Treat = 1). The time (in days) until the wound became infected was recorded; for 106 of the subjects, no infection was discovered during the period of follow up and for them total time in the study was treated as a censored survival time. Further information recorded about each subject included their Sex (Male = 0, Female = 1), Race (Non White = 0, White = 1) and Severity (the initial severity of their burns, in percent of body surface area).

(i) It was decided to fit a Cox proportional hazards model to the data, with Treat, Sex, Race and Severity 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 four explanatory variables?

(8)

	Coefficient	*Standard Error*
Treat	–0.606	0.296
Sex	–0.631	0.390
Race	2.12	1.01
Severity	0.00404	0.00703

(iii) Obtain and interpret a 95% confidence interval for the hazard ratio of a subject given additional care compared to a subject given standard care, assuming that the two subjects are of the same sex and racial class and have the same initial severity of burns.

(5)

(iv) Information was also recorded about the cause of the burns, which was characterised as either chemical (9 cases), scald (18), electric (11) or flame (116). Describe briefly how you would extend the model in order to make full use of this new information.

(2)

5. 120 patients with symptoms of hypothyroidism were carefully investigated. Each patient's T4 level on a thyroid function test was recorded. The best available methods were then used to determine whether or not each patient was truly suffering from hypothyroidism, and it was discovered that 27 of the patients were truly hypothyroid but the other 93 were not. The table shows the number of true hypothyroid and not hypothyroid patients whose T4 level (denoted by x) fell in various ranges.

	Number of Cases	
T4 value (x)	*True Hypothyroid*	*Not Hypothyroid*
$x \leq 5$	18	1
$5 < x \leq 7$	7	17
$7 < x \leq 9$	2	36
$9 < x$	0	39

A diagnostic test is to be constructed for future patients, using only the T4 level. The form of the test will be to diagnose a patient as positive for hypothyroidism if $x \leq c$, a critical value to be determined, and negative if $x > c$.

(i) Suppose the critical value, c, is set equal to 5. Estimate the sensitivity, specificity, positive predictive value and negative predictive value of the test. Referring to these values, briefly discuss the likely usefulness of this test.

(7)

(ii) Estimate the sensitivity and specificity of the test for values of $c = 7$ and $c = 9$. Draw a rough sketch of the ROC curve for this test. What value of c would you recommend? Explain why.

(10)

(iii) Using the raw data values, rather than the grouped values tabulated above, an accurate ROC curve was drawn and the area under the curve was found to be 0.86. Briefly explain what this suggests about the test.

(3)

6. (i) Define what is meant by a *case-control study*, and explain why it can be a useful research design.

(4)

(ii) Discuss one potential advantage and one potential disadvantage of a *matched* case-control study relative to an *unmatched* case-control study.

(4)

A study was carried out in a large maternity hospital to investigate the possible association between low birth weight (baby weighing less than 2500 g) and mothers smoking during pregnancy. 167 babies with low birth weight ('cases') were individually matched with babies born at the same hospital who did not have low birth weight ('controls'). The matching variable was mother's weight before pregnancy, matched to within ±100 g. Mothers were classed as either smokers or non-smokers during pregnancy. The results are displayed in the table below.

		Control Mothers	
		Smoker	*Non Smoker*
Case Mothers	*Smoker*	15	40
	Non Smoker	22	90

(iii) Calculate unmatched and matched odds ratios for the association between low birth weight and mothers smoking during pregnancy.

(4)

(iv) Carry out an unmatched chi-squared test for association and McNemar's test. Compare the results of these procedures.

(8)

7. A simple random sample of n items is chosen without replacement from a finite population of N items, and a variable, X, is measured on each unit in the sample. The values of X in the population are denoted by $X_1, X_2, \ldots, X_N$. The values that appear in the sample are denoted by $x_1, x_2, \ldots, x_n$ (where each x_i is one of the X_j). The sample mean is $\bar{x}$. In the population, X has mean $\bar{X} = \frac{1}{N}\sum_{j=1}^{N} X_j$ and variance $\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(X_j - \bar{X})^2$.

(i) Simple random sampling is defined as a process which ensures that all $\binom{N}{n}$ different possible samples have the same chance of being selected. Prove that, with simple random sampling, each item in the population has the same probability, $\frac{n}{N}$, of being selected for the sample.

(2)

(ii) Show that $\bar{x}$ is an unbiased estimator of $\bar{X}$.

(2)

(iii) The variance of $\bar{x}$ is $\mathrm{Var}(\bar{x}) = E\left\{(\bar{x} - \bar{X})^2\right\}$, where the expected value is taken over all possible samples. Prove that

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}.$$

(8)

(iv) An approximate confidence interval for $\bar{X}$ is $\left(\bar{x} - 2\sqrt{\mathrm{Var}(\bar{x})},\ \bar{x} + 2\sqrt{\mathrm{Var}(\bar{x})}\right)$. The value $B = 2\sqrt{\mathrm{Var}(\bar{x})}$ is known as the *bound on error*. Show that the sample size required to estimate $\bar{X}$ with a bound on error B is

$$n = \frac{4\sigma^2 N}{4\sigma^2 + (N-1)B^2}.$$

(4)

(v) A farmer has 1000 new-born "broiler" chickens, and intends to feed them with a new food product. When the chickens are 6 weeks old, the farmer will weigh a sample of them (weights in g) and use the resulting data to estimate the mean increase in weight of all the chickens with a bound on error of no more than 2g. The farmer believes that the population variance of weight increase is 100g^2. How big a sample of chickens is required?

(4)

8. A UK university intends to conduct a survey of its first-degree graduates from 2014, one year after their graduation, in order to find out about the jobs they are doing and their salaries. The university would like to quote the mean salary of its graduates in publicity material. It can categorise its graduates into four distinct strata, based on their subject(s) of study, the number of 2014 graduates in each stratum being known exactly from the university's own records. National data from 2013 have been used to generate the estimated means and standard deviations of salaries shown in the table below. The university believes that it will be easier to contact graduates in Stratum 4, so the costs of sampling a graduate will be cheaper in that stratum than in the other strata, as shown in the table.

			Salaries (£): 2013 Estimates	
Stratum	*Number in stratum: 2014*	*Sampling Cost*	*Mean*	*Standard Deviation*
1	1000	1.5	20 200	1550
2	1100	1.5	22 100	2150
3	1600	1.5	24 400	2950
4	600	1	29 800	1100
All	4300		23 600	3750

Throughout this question, you should ignore the finite population correction.

(i) The university is aiming for a total sample of 500 graduates. Assuming that the estimated standard deviations from 2013 are a reasonable guide to the corresponding 2014 values, find the required number of graduates in the sample from each stratum using

(a) proportional allocation,

(b) optimal allocation. (5)

(ii) Estimate the relative efficiencies of these two methods of allocation compared with a simple random sample of 500 graduates. Which sampling method would you recommend for this survey, and why? (9)

(iii) The university wishes to find out about the incomes of graduates who are in fact in employment, but does not know what proportion of its graduates are unemployed or in further study (such as MSc or PhD students). The proportion of graduates in employment will be different in the different strata. Explain how this might affect the choice of sampling method. (3)

(iv) Some graduates who are in employment are employed in graduate-level jobs, that is jobs that require them to use knowledge and skills that are usually acquired during university education, while others are employed in jobs that do not require such skills. The university has no way of knowing, in advance of carrying out this survey, how many of its graduates are in each category. It is suggested that, after collecting data from the sample, the graduates in it should be further stratified, within each of the above strata, according to whether the graduate's response suggests he or she is in a graduate-level job or another kind of job, giving 8 strata in all. Discuss this proposal. (3)

BLANK PAGE

BLANK PAGE