EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2016

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. Consider a study of the association between a binary outcome (disease versus no disease) and a binary risk factor (exposure to a risk factor versus no exposure). Let D and $\bar{D}$ denote the presence and absence of disease and E and $\bar{E}$ denote the presence and absence of exposure.

(i) Define the following concepts using conditional probabilities.

(a) Relative risk for disease.

(2)

(b) Odds ratio for a case-control study.

(2)

(c) Odds ratio for a cohort study.

(2)

(ii) Prove that the odds ratio for a case-control study equals the odds ratio for a cohort study. Explain why this result is important in medical studies of the association between a dichotomous outcome and a dichotomous risk factor.

(8)

(iii) Consider cross-sectional, case-control and cohort studies. For which of these studies can the relative risk be estimated directly? For which of these studies can the odds ratio be estimated directly?

(5)

(iv) When can the odds ratio be used as an approximation to the relative risk?

(1)

2. (i) Define the *standardised mortality ratio* (SMR). When comparing mortality in two populations, give one advantage and one disadvantage of using SMRs versus using crude death rates or age-specific rates. Give three examples of how the SMR is used in practice.

(7)

(ii) Describe *indirect standardisation*, stating explicitly how to calculate the indirectly standardised mortality rate. State any assumptions you make when applying this technique to compare populations.

(6)

(iii) When is indirect preferred to direct standardisation?

(3)

(iv) Describe the Poisson regression model for rates which underlies the technique of indirect standardisation and state clearly how to interpret this model. You may assume that the random variables of interest, e.g. numbers of deaths, have a Poisson distribution.

(4)

3. (i) Explain the main purpose of principal component analysis. Briefly discuss the decisions that need to be made when carrying out a principal component analysis. (5)

(ii) The examination marks of 88 students in five different subject areas of mathematics have been recorded. Each examination was marked out of 100 marks. Some summary statistics for these mathematics marks are given in the table below.

Examination	Mean	Variance	Standard Deviation	Minimum	Maximum
Mechanics	38.97	305.69	17.48	0	77
Vectors	50.59	172.84	13.15	9	82
Algebra	50.60	112.89	10.62	15	80
Analysis	46.68	220.38	14.85	9	70
Statistics	42.31	297.76	17.26	9	81

For these data, discuss whether it is appropriate to carry out the principal component analysis using the variance-covariance matrix or using the correlation matrix. (4)

(iii) The correlation matrix for the data is given below. Briefly describe the correlation structure in the variables. (3)

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	1.00	0.55	0.55	0.41	0.39
Vectors	0.55	1.00	0.61	0.49	0.44
Algebra	0.55	0.61	1.00	0.71	0.66
Analysis	0.41	0.49	0.71	1.00	0.61
Statistics	0.39	0.44	0.66	0.61	1.00

(iv) An analyst has extracted the principal components and eigenvalues from the correlation matrix for the data. The coefficients for the first three principal components and the eigenvalues are given in the table below.

		Component		
	Variable	1	2	3
	Mechanics	–0.40	0.65	0.62
	Vectors	–0.43	0.44	–0.71
	Algebra	–0.50	–0.13	–0.04
	Analysis	–0.46	–0.39	–0.13
	Statistics	–0.44	–0.47	0.32
Eigenvalue		3.18	0.74	0.45

(a) Interpret the first and second principal components. (2)

(b) How much of the total variation in the data is explained by the first principal component? How much of the total variation in the data is explained by the first two principal components? (2)

(c) What criteria might you use to decide on the apparent dimensionality of these data? Hence comment on the apparent dimensionality of these data. (4)

4. (i) Explain the purpose of cluster analysis and discuss briefly the decisions that need to be made when carrying out a cluster analysis.

(5)

(ii) What is the difference between a hierarchical and a non-hierarchical method of clustering? Give an example of a non-hierarchical method.

(3)

(iii) What is a dendrogram? Why is it useful?

(2)

(iv) The pairwise distances (dissimilarities) between five objects are as follows.

Object	1	2	3	4	5
1	0				
2	4	0			
3	6	9	0		
4	1	7	10	0	
5	6	3	5	8	0

Use single-linkage (also known as nearest neighbour) cluster analysis on the dissimilarity matrix above, and draw the associated dendrogram for your analysis.

(7)

(v) Discuss whether there is a clear cluster structure in the data.

(3)

5. (i) Define the *survivor function* $S(t)$ and the *hazard function* $h(t)$ for a continuous random variable T measuring lifetime. Write down an expression for the survivor function in terms of the hazard function.

(3)

(ii) The exponential distribution has constant hazard function $h(t) = \lambda$. Write down expressions for the density of the exponential distribution and the mean of this distribution in terms of λ.

(2)

(iii) Explain what is meant by a *right-censored* observation. Give two different examples of ways in which a right-censored observation might arise.

(3)

(iv) After a radical mastectomy for breast cancer, ten female patients were randomly assigned to one of two groups, an experimental group who received chemotherapy, and a control group who received no drugs. At the end of two years, survival times in months were recorded and are given in the table below. A right-censored observation is denoted by +, so 16+ denotes a right-censored observation at 16 months.

Experimental group	23	16+	18+	20+	24+
Control group	15	18	19	19	20

Compute the Kaplan-Meier estimate of the survivor function for each group and plot the results on one graph.

(10)

(v) If survival times for the control group have an exponential distribution, estimate the hazard rate.

(2)

6. Consider the following experiment on visual perception using random-dot stereograms. A random-dot stereogram is composed of two rectangles placed side by side, where each rectangle appears to consist only of randomly scattered dots, without any image. When viewed with only one eye functioning, the viewer cannot see a hidden image. However, when viewed with both eyes, if a person focuses the eyes in front of or behind the pair of images, then a three-dimensional hidden image of a diamond can be seen. Sometimes the diamond image can be seen quickly, but on other occasions it can take a while before it can be perceived. Here the response variable is the time in seconds needed to perceive the diamond image.

The experiment investigated whether giving a person prior knowledge about the shape of the image reduces the time needed to recognise it. Forty-three subjects (group NV) received just verbal information about the shape of the hidden object. Thirty-five subjects (group VV) received both verbal information and visual information, for example a drawing of the hidden object.

(i) The response time T to perceive the diamond image has a Weibull distribution with hazard function

$$h(t) = \lambda\gamma t^{\gamma-1}, \quad t \geq 0,$$

where $\lambda > 0$ and $\gamma > 0$ are parameters to be estimated. Show that the parameters are given by the intercept and slope of the theoretical relationship of the logarithm of the cumulative hazard function plotted against the logarithm of time.

(3)

(ii) Write down the proportional hazards model for a vector of p time-constant covariates, X, assuming a Weibull baseline hazard function. Interpret each term in your model. Write down an expression for the hazard ratio for this model.

(5)

(iii) Data from the random-dot stereogram experiment were analysed using a proportional hazards model with a Weibull baseline hazard function. One explanatory variable was included in the model: GroupVV, taking the value 1 if the subject was in Group VV, or 0 if the subject was in Group NV. The following edited computer output shows the results from the fitted model.

	Estimate	Standard Error
λ	0.060	0.019
γ	1.260	0.104
GroupVV	0.552	0.233

(a) Use the fitted model to estimate the hazard ratio for a subject in the VV Group compared to a subject in the NV Group. Construct a 95% confidence interval for this hazard ratio. Which group, on average, has the shorter response times?

(3)

Question 6 continued on the next page

(b) What is the estimated hazard function for a subject in Group NV? What is the estimated hazard function for a subject in Group VV? How does the estimated hazard function change with time for each group?

(3)

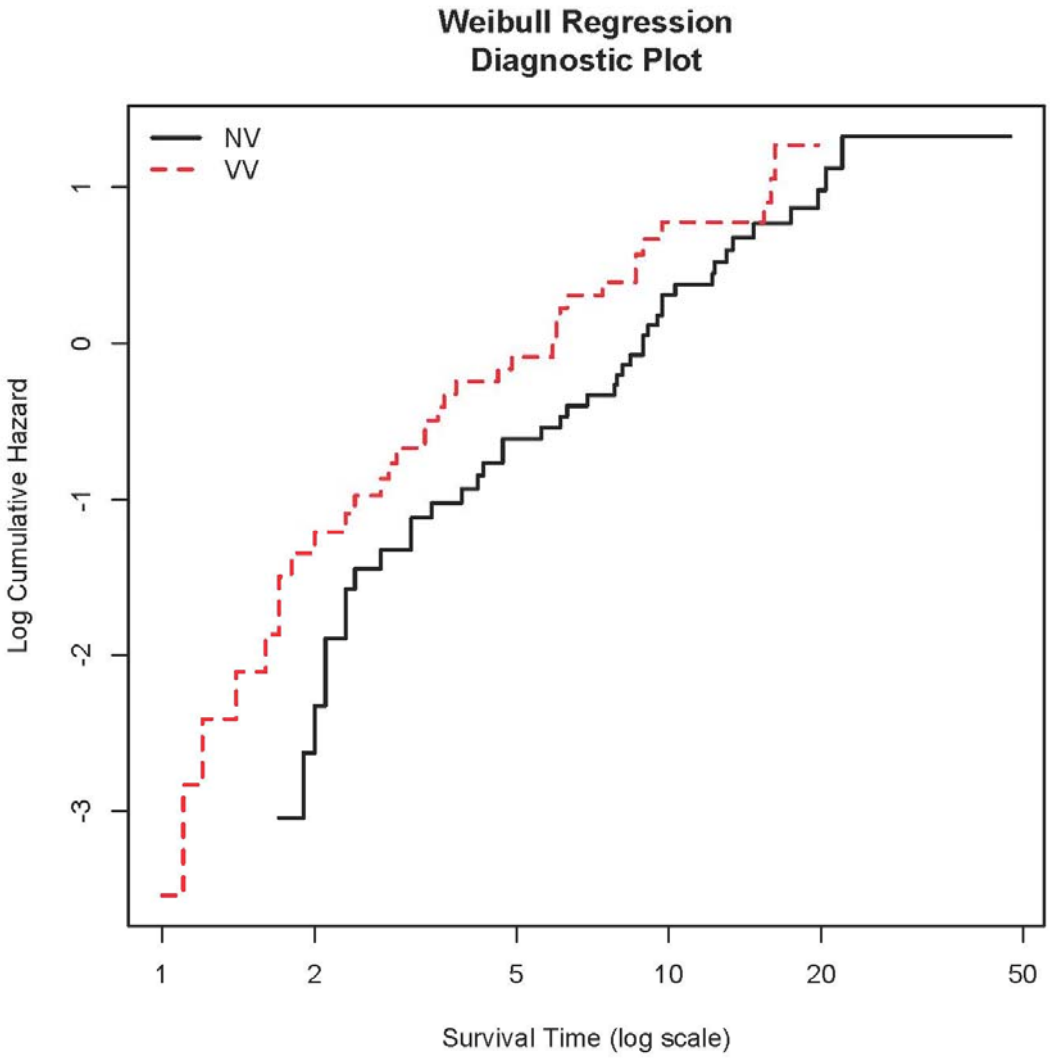
(iv) Someone suggested using an exponential distribution, instead of a Weibull distribution, to model the response times. What advice should you give regarding this idea? Justify your answer.

(3)

(v) The graph below shows estimates of the log cumulative hazard function for the NV group (solid line) and the VV group (dashed line). Referring to this graph, discuss whether the Weibull proportional hazards model is appropriate for the stereogram response times. Justify your answer.

(3)

$\log(-\log(\hat{S}(t)))$ for Group NV (solid line) and Group VV (dashed line)



7. (i) For a simple random sample without replacement from a finite population, the sample mean is an unbiased estimator of the population mean and has variance

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 .$$

Define the symbols in the above formula.

(4)

(ii) Use the formula in part (i) to prove that

$$V(\hat{p}) = \frac{N-n}{N-1} \times \frac{P(1-P)}{n}$$

where $\hat{p}$ and P denote the sample and population proportion of a certain characteristic.

(3)

There are 650 Members of Parliament (MPs) in the United Kingdom (UK) House of Commons. MPs can employ staff to help run their offices. MPs may employ as many office staff as they wish, subject to a fixed budget. A 20% simple random sample was selected in order to estimate the total number of office staff currently employed by MPs. The following data on the number of staff employed were collected.

Number of employees	1	2	3
Frequency	25	80	25

(iii) Estimate the total number of office staff currently employed by MPs and use the result in part (i) to construct an approximate 95% confidence interval for this total.

(5)

Shingles is a painful skin rash caused by the reactivation of the chickenpox virus in people who have previously had chickenpox. A shingles vaccine is recommended by the UK's National Health Service to people aged 70 or older.

(iv) On 30 April 2012, there were 405 members of the UK's House of Lords aged 70 or older. A researcher plans to take a simple random sample of these 405 people in order to estimate the percentage who have been vaccinated against shingles to within a margin of error of 1%. The researcher asks you to estimate the smallest achieved sample size that would be necessary to do so with 95% confidence. Using the result in part (ii), make this estimate. What advice should you give the researcher regarding the use of the smallest sample size that you found and regarding whether their over-age-70 Lordships are a suitable source of a sample that is in some sense representative of all UK adults over 70?

(8)

8. (i) Define the term *stratified random sampling*.

(2)

(ii) Explain the circumstances when stratified random sampling may be expected to work well, and give an example of such a situation.

(2)

(iii) Explain in words what is meant by the expressions *stratification with proportional allocation*, *stratification with Neyman allocation* and *stratification with optimal allocation*. In your answer, you should explain in what sense each of "Neyman allocation" and "optimal allocation" makes the most effective use of resources. Give three reasons why these allocations are only theoretical.

(11)

(iv) State the optimal allocation formula for the allocation of sample size in stratified random sampling. Define all the symbols used in the formula.

(5)

BLANK PAGE

BLANK PAGE

BLANK PAGE