EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2013

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. An engineer wished to investigate the effect of an additive in petrol. He was interested in efficiency defined by 'miles per gallon'. He chose 20 cars, and randomly selected 10 to have the additive. He filled each car with petrol, adding the extra additive in the random selection, and ran them on a track, recording their mileage per gallon. He then repeated the procedure on a different track, again recording the mileage per gallon. Drivers did not know whether the car they were driving had the additive, and they were all instructed to drive the cars at a given speed.

Mileage per gallon is recorded in the variable `eff1` for data from the first track and in the variable `eff2` for data from the second track.

Summary statistics are given below.

```
-------------------------------------------------
-> additive = no

    Variable |       Obs        Mean    Std. Dev.
-------------+--------------------------------
        eff1 |        10       21.35    3.0351
        eff2 |        10       20.22    2.6824
correlation of eff1 and eff2 is 0.9907
-------------------------------------------------
-------------------------------------------------
-> additive = yes

    Variable |       Obs        Mean    Std. Dev.
-------------+--------------------------------
        eff1 |        10       22.79    3.4806
        eff2 |        10       22.01    3.3828
correlation of eff1 and eff2 is 0.9799
-------------------------------------------------
```

(i) Explain why Hotelling's T^2 test could be used to test the effect of the additive, and state the null and alternative hypotheses.

(4)

(ii) Use the summary statistics to compute Hotelling's T^2 statistic.

(10)

(iii) Use your result from part (ii) to compute an F value and carry out a suitable test of the hypotheses in part (i).

(3)

(iv) What are your conclusions from this test?

(3)

2. A market researcher has collected data from customers of a mail order company. She has responses from 5000 customers including the following variables.

- Age in years
- Number of years they have been customers
- Sex
- A score of social deprivation (scored 1 to 9 and derived from the area in which the customer lives)
- Amount of money spent with the company in the past year
- 10 answers rating satisfaction with different aspects of the company scored on a scale 1–5 where 1 indicates high dissatisfaction and 5 indicates high satisfaction

The researcher wants to carry out suitable multivariate analyses to answer the following questions.

A. What are the main patterns in the satisfaction data?

B. Do customers belong to identifiable subgroups in relation to their demographics?

C. Do the characteristics of customers who spent more than £500 in the last year differ from those who spent up to £500 in the last year?

(i) For each of the 3 questions provide the following information.

- A multivariate method that could be used to answer the question
- Any coding of the data that may be necessary and why
- Other decisions that would need to be made about the analysis

(15)

(ii) Two people have independently tried to answer B, and have identified different groupings. Assuming that they have each used an appropriate method, how would you explain this apparent problem to the researcher?

(5)

3. (i) State what is meant by *censored data*, and explain why they often occur in survival analysis. (3)

A doctor is investigating the risk of patients contracting disease D after an invasive procedure P. He has data on 20 patients, 8 of whom were known to have contracted D after having P. He has the time in months, since P, until either the disease was contracted or the patient was last known to be disease-free.

The times, in months, are given below, where * indicates that the disease was not contracted in that period of follow-up.

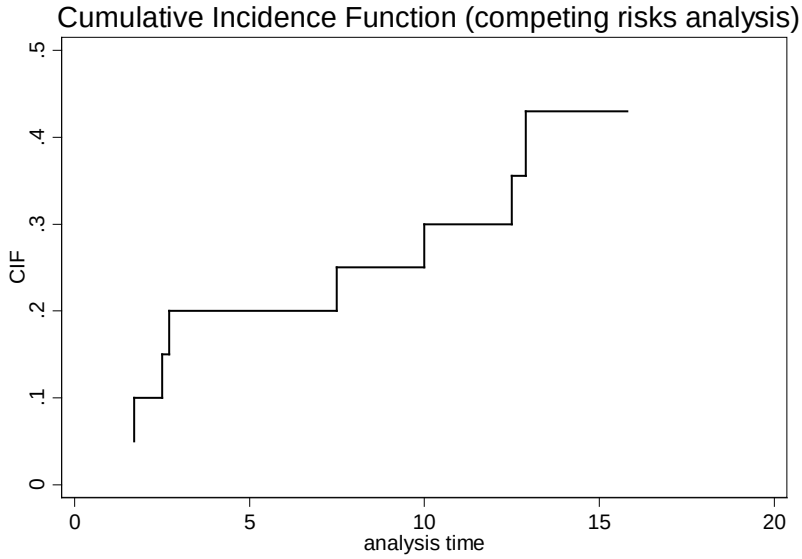
1.7	1.7*	2.5	2.7	3.4*	4.8*	7.5	9.7*	10.0	12.5
12.7*	12.9	13.7*	14.6*	15.6*	15.8	16.9*	17.3*	17.9*	18.3*

(ii) Calculate the Kaplan-Meier survivor function for these data. (5)

(iii) The doctor has been advised to study the cumulative incidence function $C(t)$ rather than the survivor function. This is given by $C(t) = 1 - S(t)$. Plot a graph of the cumulative incidence function derived from your results in part (ii), and interpret it. (4)

(iv) After doing this analysis you are told that some of the patients who did not contract D actually died, that some deaths were related to P, and that the corresponding recorded censored times are the times from P to their death. Explain why your Kaplan-Meier estimates are now invalid. (3)

(v) A statistician uses a procedure known as 'competing risks' to produce a cumulative incidence function for D which takes deaths into account. This is shown below.



(a) Explain why the risk 'death' competes with the risk of contracting D. (2)

(b) Compare this function with your graph in part (iii). Explain why the curves are different. (3)

4. The Weibull distribution for a random variable T has probability density function

$$f(t) = \lambda\gamma t^{\gamma-1}e^{-\lambda t^{\gamma}} \quad t \geq 0, \gamma > 0, \lambda > 0,$$
$$f(t) = 0 \quad \text{otherwise.}$$

(i) State how the exponential distribution is related to the Weibull distribution. (1)

(ii) Derive the survivor function $S(t)$ for the Weibull distribution. (2)

(iii) Derive the hazard function, as used in survival analysis, for the Weibull distribution. (2)

(iv) Show that, if T has an exponential distribution, a plot of $\log(-\log(S(t)))$ against $\log(t)$ is linear, and state the intercept and gradient of the line. (3)

(v) An engineer is interested in how the level, X, of a chemical affects the lifetimes of a certain type of component. He has survival data on 17 components, describing the level of the chemical and the lifetime of the component.

(a) Parametric regression models have been fitted using the exponential and Weibull distributions. The outputs are given below. Interpret the outputs, stating which model you would prefer and why. (5)

Exponential regression -- log relative-hazard form

```
No. of subjects =         17                        Number of obs =     17
No. of failures =         17
Time at risk    =       1012
                                                    LR chi2(1)    =   6.49
Log likelihood  =  -27.42037                        Prob > chi2   = 0.0108
------------------------------------------------------------------------------
             | Estimate   Std. Err.     z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Haz. Ratio x | 2.762984   1.088195    2.58   0.010     1.276835    5.978911
------------------------------------------------------------------------------
```

Weibull regression -- log relative-hazard form

```
No. of subjects =         17                        Number of obs =     17
No. of failures =         17
Time at risk    =       1012
                                                    LR chi2(1)    =   5.96
Log likelihood  = -27.413446                        Prob > chi2   = 0.0146
------------------------------------------------------------------------------
             | Estimate   Std. Err.     z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Haz. Ratio x | 2.730445    1.11054    2.47   0.014     1.23035      6.05952
-------------+----------------------------------------------------------------
       Gamma | .9767104   .1964255                      .6585386    1.448606
------------------------------------------------------------------------------
```

Question 4 continues on the next page

(b) The output from a Cox proportional hazards model is given below. State how this model is different from the parametric models, interpret this output and contrast the results with those from part (a).

(4)

Cox PH regression

```
No. of subjects =         17                        Number of obs =     17
No. of failures =         17
Time at risk    =       1012
                                                    LR chi2(1)    =   9.68
Log likelihood  = -28.858353                        Prob > chi2   = 0.0019
------------------------------------------------------------------------------
             | Estimate   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Haz. Ratio x | 4.170992   2.000554    2.98   0.003     1.629196    10.67838
------------------------------------------------------------------------------
```

(c) It is known that, if a survival model fits well, the residuals known as Cox-Snell residuals have an exponential distribution with $\lambda = 1$. Briefly describe how you could use Cox-Snell residuals and your result from part (iv) to investigate which of the models best describes the data.

(3)

5. (i) Define the term *confounding factor* in the context of epidemiology.

(1)

(ii) Explain why it is difficult to infer causality from observational studies. Illustrate your answer with two scenarios: one where cause and effect is not credible, and another where it is credible.

(4)

(iii) A medical doctor from your local hospital has written to you asking for advice on the design of a study she wants to carry out. An extract from the letter is given below.

> "Procedure X is carried out on some pregnant mothers during labour. I believe that X increases the incidence of adverse events for the mother and/or the baby, and want to run a study to prove this. There are several possible adverse events, each of which is quite rare (e.g. maternal death). So I want to study a composite outcome where the outcome is bad if any one of a set of adverse events happens to the mother or the baby; otherwise it is good. Each of these events happens at or soon after delivery of the baby. X is carried out on about 300 mothers per year and data about X and about the adverse events are routinely recorded on the hospital database. Can you please advise me on the appropriate study design?"

Draft a letter of reply to the doctor. You should use language which is understandable by the doctor who can be assumed to have little statistical expertise. You should give your advice and outline your reasons and also any other information you will need to make a more definite recommendation.

(15)

6. (i) Define what is meant by a *case-control study*, outlining the situations in which it is a useful research design.

(4)

(ii) Distinguish between a matched and an unmatched case-control study.

(2)

A researcher has collected data from a matched case-control study, studying the association between exposure to a potential risk factor and the subsequent development of a disease. The results are given below.

		Controls	
		Exposed	Not exposed
Cases	Exposed	12	8
	Not exposed	10	7

The researcher has done an unmatched analysis, but she has been told that this is incorrect. She argues that the results are similar for the matched and unmatched analyses.

(iii) Carry out an unmatched chi-squared test for association and McNemar's test, and compare their results.

(8)

(iv) Compute unmatched and matched odds ratios for the association between risk factor and disease for these data, and compare them.

(2)

(v) Explain why it is desirable to carry out a matched analysis for matched data.

(4)

7. You have been asked to analyse data from a survey of residents in a particular city, where the unit of enquiry is resident. The table below gives summary statistics for a continuous variable *V* measured in the survey. Unfortunately some of the details of the survey have gone missing and, in particular, it is not known how the three groups in the city were defined.

Group	*Size of group*	*Size of sample*	*Mean*	*Variance*
1	1000	100	15	3
2	2000	100	18	8
3	3000	100	23	10

(i) Define three groups in a city for which this would not be a true stratified sample, giving your reasons.

(4)

(ii) Describe a different scenario where this would be a stratified sample, and where such a design would be appropriate. Justify your answer.

(4)

(iii) Assuming that the sample is a valid stratified sample, calculate an estimate and approximate 95% confidence interval for the population mean of *V*.

(5)

(iv) You are to design a sample survey measuring *V* in a different city. For practical reasons you are restricted to a sample of size 300, and you are required to sample from the three strata as defined in the table. The relative sizes of the strata in the two cities are similar, but the stratum variances may differ from one city to the other. Discuss two possible methods for allocating the 300 sampling units over the strata, and how you might decide on the one to recommend. Illustrate your allocations by applying them to the data for the city in the table above, commenting briefly on the outcomes.

(7)

8. (i) Define a *cluster* as used in sampling. (1)

(ii) What is the *intra-cluster correlation* (ICC) and why is it an important statistic in sampling theory? (2)

(iii) Describe **either** a scenario where you would expect the ICC to be low **or** a scenario where you would expect it to be high. State the population, the outcome, and the clusters. Justify your answer. (4)

(iv) Outline the merits and disadvantages of cluster sampling. (4)

(v) It is known that if a population of MN units comprises M clusters, each of size N, and if m of these clusters are chosen and measured, the variance of the cluster sampling mean is approximately

$$\text{Var}(C) = \text{Var}(R)(1 + (N - 1)\rho)$$

where $\text{Var}(R)$ is the variance of the simple random sampling mean for a sample of the same size (mN) ignoring the cluster structure and ρ is the intra-cluster correlation.

A population is divided into 100 clusters each of size 20 and a survey is carried out. A simple random sample of 10 clusters is selected and a characteristic Y is measured on each member of the 10 clusters.

(a) In order to improve the precision of the results it is suggested that the clusters are halved in size and twice as many are selected, so that the total number of sampled items remains the same.

Assuming that the ICC remains the same, write down an expression for the ratio of the variances from the two sampling schemes, and by means of a graph show how this ratio depends on the ICC in the range 0 to 0.3. (2)

(b) Comment on the validity of the assumption that ICC is the same for smaller clusters. (1)

(c) The means for the ten clusters are

12.6 18.4 13.3 9.5 21.1 11.1 15.9 10.3 12.0 14.4,

with $\sum_{i=1}^{10}\sum_{j=1}^{20} y_{ij} = 2772$ and $\sum_{i=1}^{10}\sum_{j=1}^{20} y_{ij}^2 = 42\,200$ where y_{ij} is the value of Y for the jth person in cluster i.

Compute an estimate and approximate 95% confidence interval for the mean of the population using cluster sampling methodology. Deduce the ICC and use this to estimate the width of the 95% confidence interval which would be expected if the alternative sampling scheme, using 20 clusters, was used. (6)

BLANK PAGE

BLANK PAGE