# EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

**GRADUATE DIPLOMA, 2012**

**MODULE 2 : Statistical inference**

**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 7 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. The random variables $X_1$, $X_2$, …, $X_n$ represent a random sample from the Poisson distribution with unknown mean $\lambda > 0$. An estimator of $e^{-\lambda}$ is to be sought.

(i) Show that $\sum_{i=1}^{n} X_i$ is a sufficient statistic for $\lambda$.

(4)

(ii) Show that the Cramér-Rao lower bound for the variance of unbiased estimators of $e^{-\lambda}$ is $\frac{\lambda e^{-2\lambda}}{n}$.

(5)

(iii) For $i = 1, 2, \ldots, n$, the random variable $Y_i$ takes the value 1 when $X_i = 0$ and the value 0 otherwise. Show that the sample mean $\bar{Y}$ is an unbiased estimator of $e^{-\lambda}$ and find its efficiency.

(6)

(iv) By first considering $E\left\{ 1-\frac{1}{n}^{X_i} \right\}$, show that $1-\frac{1}{n}^{\Sigma X_i}$ is also an unbiased estimator of $e^{-\lambda}$. Say, with reasons but without deriving any further results, which of $\bar{Y}$ and $1-\frac{1}{n}^{\Sigma X_i}$ you might prefer as an estimator of $e^{-\lambda}$.

(5)

2. In an experiment, the proportion $X$ of cells infected by a virus is randomly distributed with probability density function $\frac{\alpha(1+x)^{-(1+\alpha)}}{1-2^{-\alpha}}$ for $0 < x < 1$, where $\alpha > 1$ is an unknown parameter. The proportions, $X_1$, $X_2$, …, $X_n$ , have been obtained in $n$ independent experiments.

(i) Show that $\hat{\alpha}$ , the maximum likelihood estimator of $\alpha$, satisfies the following equation.

$$\frac{n}{\hat{\alpha}} - \frac{n\log 2}{2^{\hat{\alpha}} - 1} - \sum_{i=1}^{n} \log(1 + X_i) = 0$$

[Do not attempt to solve this equation.]

(6)

(ii) In the case $n = 100$ and $\sum \log(1 + x_i) = 20$, an initial estimate of $\hat{\alpha}$ is 4.0. Use one iteration of the Newton-Raphson method to find a better estimate.

(8)

(iii) Using the estimate found in part (ii), find an approximate 95% confidence interval for $\alpha$.

(6)



**Turn over**

3. A device measures the distances between minor flaws along a fibre optic cable. The flaws occur at random at average rate $\lambda > 0$ (known). However, flaws within an unknown distance $\delta > 0$ of each other cannot be distinguished by the device. The measured distance $x$ between flaws has probability density

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-\delta)} & x > \delta, \\ 0 & \text{otherwise.} \end{cases}$$

A random sample of distances $X_1, X_2, \ldots, X_n$ is available and $Y = \min(X_1, X_2, \ldots, X_n)$.

(i) By sketching a plot of the likelihood, show that $Y$ is the maximum likelihood estimate of $\delta$.

(5)

(ii) Show that $P(Y \geq y) = \exp(-n\lambda(y - \delta))$ for $y > \delta$ and hence find the distribution function of $Y - \delta$.

(5)

(iii) Define a *pivotal quantity*. Show that $Y - \delta$ is a pivotal quantity for $\delta$.

(5)

(iv) Find the value of $c$ such that $[Y - c, Y]$ is a 95% confidence interval for $\delta$.

(5)



**Turn over**

4. Survival times in Populations A and B have probability density functions $f_A(x) = \alpha^2 x e^{-\alpha x}$ and $f_B(y) = \beta^2 y e^{-\beta y}$ respectively, where $\alpha > 0$ and $\beta > 0$ are unknown parameters. A random sample of size $n_1$, $X_1, X_2, \ldots, X_{n_1}$, is available from Population A and an independent random sample of size $n_2$, $Y_1, Y_2, \ldots, Y_{n_2}$, is available from Population B. It is required to test the null hypothesis $\alpha = \beta$ against the alternative hypothesis $\alpha \neq \beta$.

(i) Show that the generalised likelihood ratio test has critical region of the form

$$n_1 \log \frac{n_1}{n_1 + n_2}(1 + W) + n_2 \log \frac{n_2}{n_1 + n_2}\left(1 + W^{-1}\right) \geq k,$$

for some constant $k$, where $W = \frac{\sum Y_i}{\sum X_i}$.

(9)

(ii) Carry out this test at the 5% level in the case $n_1 = 100$, $n_2 = 200$, $\Sigma x_i = 800$ and $\Sigma y_i = 1500$ and report your conclusions.

(6)

(iii) In the general case, explain very briefly how the generalised likelihood ratio test would be modified if it were required to test the null hypothesis $\alpha = \beta + \Delta$ against $\alpha \neq \beta + \Delta$ at the 5% level, where $\Delta$ is a known constant. Also, explain how an approximate 95% confidence interval for $\alpha - \beta$ could be deduced from this type of test.

(5)

5. (a) Use the one-sample Kolmogorov-Smirnov test to examine whether the following data appear to come from a uniform distribution between 0 and 2.

1.3 0.7 1.0 1.5 1.9 1.7 1.4 0.6 0.9 0.8

[You are given that the critical value for the test statistic at the 5% significance level is 0.409.]

(7)

(b) The independent random variables $X_1$, $X_2$, $X_3$ have a common continuous distribution with median $\theta$. It is required to find a 95% confidence interval for $\theta$ using the bootstrap method based on the sample median.

(i) Explain the method for constructing such a confidence interval.

(3)

(ii) Find the probability that $\theta < \min(X_1, X_2, X_3)$ and the probability that $\theta > \max(X_1, X_2, X_3)$. What implication does your answer have for the usefulness of the bootstrap method in this case?

(4)

(c) A random sample $Y_1, Y_2, \ldots, Y_n$ is available from a probability distribution with standard deviation $\sigma$ and it is required to estimate $\sigma$. The biased estimator $\hat{\sigma} = \sqrt{\frac{\sum Y_i - \bar{Y}^2}{n-1}}$ has been proposed, where $\bar{Y}$ is the sample mean.

(i) Describe how the jack-knife estimator based on $\hat{\sigma}$ is found. In what respect would you expect this estimator to be preferred to $\hat{\sigma}$?

(3)

(ii) Explain how an approximate 95% confidence interval for $\sigma$ may be found based on the jack-knife method.

(3)

6. A colleague has used statistical methods in the analysis of data, but has been told that he is using a frequentist approach whereas a Bayesian method might be better. He asks you to describe how the approaches differ, and how easy the Bayesian approach is to use. What would be your reply?

(8)

A doctor is investigating the probability, $p$ ($0 < p < 1$), of recovery within 6 months from a certain medical condition. Her prior distribution for $p$ is beta, with parameters $\alpha = 12.0$ and $\beta = 6.0$. She follows up a random sample of 60 patients with this condition and finds that 45 recover within 6 months.

[The beta distribution with parameters $\alpha$ ( > 0) and $\beta$ ( > 0) has probability density function

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad (0 < x < 1),$$

where $\Gamma(.)$ is the gamma function, and may be approximated by the Normal distribution with mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$ when $\alpha$ and $\beta$ are large.]

(i) Find the posterior distribution of $p$.

(4)

(ii) Find an approximate 95% Bayesian interval for $p$.

(4)

(iii) Four new patients have just been diagnosed as having the condition. Assuming that their outcomes are independent, find the probability that 3 of them recover within 6 months for a given value of $p$. Use the posterior distribution in part (i) to find the predicted value of this probability.

(4)

7. Explain what is meant by an *inadmissible* decision rule.

(2)

Three diagnostic tests, A, B, C, for a certain disease are available. Each test has only two outcomes: positive or negative. For a patient with the disease, Tests A, B and C will be positive with probabilities 0.9, 0.7 and 0.8 respectively, while for a patient who does not have the disease, Tests A, B and C will be negative with probabilities 0.5, 0.9 and 0.8 respectively. Tests A and B cost 2 units per patient, while Test C costs only 1 unit. Also, a positive test for a patient without the disease incurs an additional cost of 10 units, while a negative test for a patient with the disease incurs an additional cost of 40 units. (There is no additional cost for a correct diagnosis.) Only one diagnostic test can be used.

(i) Evaluate the risk function for each of the three decisions (i.e. choice of test), and draw a suitable diagram. You may use this diagram when answering the following three parts.

(6)

(ii) State which, if any, of the decisions are inadmissible.

(3)

(iii) Find the randomised decision rule which is minimax.

(5)

(iv) What is the Bayes solution if the prior probability that a patient has the disease is 0.1?

(4)

8. Describe the uses of the *Central Limit Theorem* in statistics. In what ways can the statistician check the reasonableness of an approximation or assumption that is made when appealing to the Central Limit Theorem?

(20)