

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



GRADUATE DIPLOMA, 2002

Statistical Theory and Methods II

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 silent, cordless, non-programmable electronic calculators.*

*Where a calculator is used the **method** of calculation should be stated in full.*

*Note that  $\binom{n}{r}$  is the same as  ${}^n C_r$  and that  $\ln$  stands for  $\log_e$ .*



1. It is believed that the number of breakages in a damaged chromosome,  $X$ , follows a truncated Poisson distribution with probability mass function

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{1 - e^{-\lambda}} \frac{1}{k!}, \quad k = 1, 2, \dots,$$

where  $\lambda > 0$  is an unknown parameter. The frequency distribution of the number of breakages in a random sample of 33 damaged chromosomes was as follows.

<i>Number of breakages</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	<i>Total</i>
<i>Number of chromosomes</i>	11	6	4	5	0	1	0	2	1	0	1	1	1	33

- (i) Assuming that the appropriate regularity conditions hold, find an equation satisfied by  $\hat{\lambda}$ , the maximum likelihood estimate of  $\lambda$ . (6)
- (ii) Assuming that an initial guess of  $\hat{\lambda}$  is available, use the Newton-Raphson method to find an iterative algorithm for computing the value of  $\hat{\lambda}$ . There is no need to carry out any numerical calculations. (6)
- (iii) It is found that the observed data give the estimate  $\hat{\lambda} = 3.6$ . Using this value for  $\hat{\lambda}$ , test the null hypothesis that the number of breakages in a damaged chromosome follows a truncated Poisson distribution. The categories 6 to 13 breakages should be combined into a single category in the goodness-of-fit test. (8)

2. Define a *loss function* and the *Bayes risk* in the context of decision theory. Also explain what is meant by a *conjugate family of distributions*.

(5)

The probability that a certain tomato seed germinates is  $\theta$ . A gardener sows a set of  $n$  such seeds and finds that  $x$  of them germinate. It can be assumed that the seeds germinate independently of one another.

- (i) Find the posterior distribution of  $\theta$ , assuming that its prior distribution is beta with parameters  $\alpha$  and  $\beta$ .

(5)

- (ii) Assuming a quadratic loss function and stating clearly any result that you use, find the Bayes estimate of  $\theta$ .

(3)

- (iii) Suppose that the gardener can make one of two decisions,  $d_0$  or  $d_1$ . The decision  $d_0$  is taken when  $\theta$  is small and the loss is then  $c\theta^2$  (where  $c > 0$ );  $d_1$  is taken when  $\theta$  is large and there is then a unit loss. Show that the Bayes risk is minimised if the gardener decides that  $\theta$  is large if

$$\frac{\alpha + x}{\alpha + \beta + n} > \frac{1}{c} \frac{\alpha + \beta + n + 1}{\alpha + x + 1}.$$

What is the gardener's decision if a uniform prior distribution is used for  $\theta$ ,  $n = 15$ ,  $x = 10$  and  $c = 25$ ?

(7)

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

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

where  $\Gamma(\cdot)$  denotes the gamma function.]

3 (i) Define the *Spearman rank correlation coefficient* and explain how it is related to the product moment correlation coefficient. You may assume that scores are not tied.

(5)

(ii) Seven students are each assessed by two examinations, *A* and *B*. It is required to test the null hypothesis that there is zero correlation between the marks for the two examinations against the alternative hypothesis that there is a positive correlation. The rankings for the *A* marks for the seven students are 1, 2, 3, 4, 5, 6 and 7, respectively, while the corresponding rankings for the *B* marks are 1, 2, 3, 4, 5, 7 and 6. *Without* using tables, show that the *p*-value for these data is  $1/720$ .

(7)

(iii) A biologist is interested in whether there is any association between the sizes of two organisms *X* and *Y* grown in different environments. Data for their sizes, in cubic centimetres, after a fixed time for eight different environments were as follows.

<b>Environment</b>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>
<i>Organism X</i>	22	16	38	187	24	68	31	478
<i>Organism Y</i>	44	48	32	155	42	93	35	336

Use Spearman's rank correlation test to test the hypothesis of no association between the sizes of the two organisms. [You may use Table XVI of the "Abridged Tables for use by Examination Candidates" *in part (iii)*.]

(8)

4. Explain what is meant by the *power* of a test and describe how the power may be used to help determine the most appropriate size of a sample. (4)

Let  $X_1, X_2, \dots, X_n$  be a random sample from the Weibull distribution with probability density function

$$f(x) = \theta \lambda x^{\lambda-1} \exp(-\theta x^\lambda), \quad x > 0,$$

where  $\theta > 0$  is unknown and  $\lambda > 0$  is known.

- (i) Find the form of the most powerful test of the null hypothesis that  $\theta = \theta_0$  against the alternative hypothesis that  $\theta = \theta_1$ , where  $\theta_0 > \theta_1$ . (6)
- (ii) Find the distribution function of  $X^\lambda$  and deduce that this random variable has an exponential distribution. (3)
- (iii) Use  $\chi^2$  tables to find the critical region of the most powerful test at the 1% level when  $n = 50$ ,  $\theta_0 = 0.05$  and  $\theta_1 = 0.025$ . (4)
- (iv) Evaluate the power of the test found in part (iii). (3)

[If  $Y_1, Y_2, \dots, Y_m$  is a random sample from an exponential distribution with mean  $v^{-1}$ , then  $2v \sum_{i=1}^m Y_i$  has a  $\chi_{2m}^2$  distribution.]

5. Explain what is meant by a *maximum likelihood estimator* and state the large-sample properties of this type of estimator under standard regularity conditions.

(5)

Suppose that the numbers of a particular plant species in sampling quadrats follow a Poisson distribution with mean  $\lambda$  and that it is required to estimate  $\theta = \lambda^2$ . A random sample of  $n$  such quadrats yields the numbers  $X_1, X_2, \dots, X_n$ .

- (i) Show that  $\hat{\theta} = \bar{X}^2 - \frac{\bar{X}}{n}$ , where  $\bar{X}$  denotes the sample mean, is an unbiased estimator of  $\theta$ .

[Standard results concerning the Poisson distribution may be assumed.]

(4)

- (ii) Show that the Cramér-Rao lower bound for the variance of unbiased estimators of  $\theta$  is

$$\frac{4\theta^{3/2}}{n}.$$

(5)

- (iii) When  $n = 1$ , show that

$$E[\hat{\theta}^2] = \lambda^2 E[(X+1)(X+2)],$$

and hence that

$$\text{Var}(\hat{\theta}) = 2\lambda^2(2\lambda + 1).$$

Use these results to find the efficiency of  $\hat{\theta}$  when  $n = 1$ , and comment on the value of the efficiency when  $\theta$  is large.

(6)

6. Explain carefully the relationship between *statistical tests* and *confidence sets* in classical statistical inference. (6)

Suppose that  $X_1, X_2, \dots, X_n$  is a random sample from the shifted exponential distribution with probability density function

$$f(x) = \frac{1}{\theta} e^{-(x-\mu)/\theta}, \quad \mu < x < \infty,$$

where  $\theta > 0$  and  $-\infty < \mu < \infty$ . Both  $\theta$  and  $\mu$  are unknown, and  $n > 1$ .

- (i) The sample range  $W$  is defined as  $W = X_{(n)} - X_{(1)}$ , where  $X_{(n)} = \max_i X_i$  and  $X_{(1)} = \min_i X_i$ . It can be shown that the joint probability density function of  $X_{(1)}$  and  $W$  is given by

$$f_{X_{(1)}, W}(x_{(1)}, w) = n(n-1)\theta^{-2} e^{-n(x_{(1)}-\mu)/\theta} e^{-w/\theta} (1 - e^{-w/\theta})^{n-2},$$

for  $x_{(1)} > \mu, w > 0$ .

Hence obtain the marginal probability density function of  $W$  and show that  $W$  has distribution function

$$P(W \leq w) = (1 - e^{-w/\theta})^{n-1}, \quad w > 0. \quad (6)$$

- (ii) Show that  $W/\theta$  is a pivotal quantity. (4)
- (iii) Without carrying out any calculations, explain how the result in part (ii) may be used to construct a  $100(1 - \alpha)\%$  confidence interval for  $\theta$  for  $0 < \alpha < 1$ . (4)

7. Give an account of the classical, Bayesian and likelihood approaches to hypothesis testing. Your answer should include, for example, details of how the tests are constructed in each case, and discussion of the merits and drawbacks of the three approaches.

(20)

8. A system is either active or under repair. The duration of active times,  $X$ , and the duration of repair times,  $Y$ , are independently exponentially distributed with respective unknown means  $\theta$  and  $\phi$ . System availability is defined as  $\psi = \theta / (\theta + \phi)$  and it is required to test the null hypothesis that  $\psi = 0.5$  against the alternative that  $\psi = 0.7$  with approximate Type I and II errors of 0.05. The sequence of paired observations  $\{(x_i, y_i), i = 1, 2, \dots\}$  is available.

(i) Write down the joint probability density function of  $X$  and  $Y$ , and hence show that  $P(Y \leq X) = \theta / (\theta + \phi)$ .

(5)

(ii) Derive a sequential probability ratio test of the above hypotheses based on the binomial distribution, using the sequence of observations

$$w_i = \begin{cases} 1, & y_i \leq x_i, \\ 0, & y_i > x_i \end{cases}$$

for  $i = 1, 2, \dots$

(7)

(iii) Construct a graph to show how the test in part (ii) may be carried out.

(3)

(iv) Find the approximate expected sample size for the above test under the alternative hypothesis.

(5)