EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2004

Paper I : Statistical Theory

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.*

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 9 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. Of the adult population in a large city, 60% favour a new leisure complex, 30% oppose it and 10% are indifferent. A random sample of 4 adults is taken from the population and their opinions on the new complex are noted.

(i) Find the probability that

(a) all four think alike,

(b) none of the four is opposed to the complex,

(c) all three opinions (in favour, oppose, indifferent) are represented in the sample,

(d) all four are in favour of the complex, if it is given that none of the four is opposed.

(10)

(ii) State the expectation and variance of the number in the sample who are in favour of the complex.

(3)

(iii) In this city, one quarter of adults are classified as "young" (age < 30) and three-quarters are "older" (aged at least 30). You are told that 12% of young adults oppose the complex; deduce the proportion of older adults who are opposed.

(3)

(iv) Given that the sample consists of one young adult and three older adults, find the probability that exactly one member of the sample opposes the complex.

(4)

Turn over

2. (i) The random variable X follows the Poisson distribution with probability mass function

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots .$$

(a) Find the moment generating function of X.

(b) Hence or otherwise show that the mean and variance of X are both equal to λ.

(c) State the Poisson approximation to the binomial distribution, indicating the circumstances in which it is appropriate.

(8)

(ii) A civil servant calculates weekly social security payments for unemployed adults. These payments vary according to claimants' circumstances, and errors may occur. Over a long period of time, the probability of a wrong calculation has been found to be 0.0075. Find to 4 decimal places the exact probability that a sample of 200 contains (a) 1 wrong calculation, (b) 4 wrong calculations.

(5)

(iii) Repeat part (ii) using the Poisson approximation to the binomial distribution. In each case find to two significant figures the percentage error in the Poisson calculation. Comment briefly on your results.

(7)

3. Apple juice is dispensed by a machine into cartons. The nominal volume of apple juice in a carton is 1 litre (1000 ml). The actual volumes of juice put into the cartons can be regarded as being independently Normally distributed, with mean set at 1010 ml and standard deviation 8 ml.

(i) Find the proportion, in a long run of production, of cartons containing less than the nominal volume.

(4)

(ii) Cartons of apple juice are often sold in packs of 6. Write down the distribution of the total volume of juice in a pack of 6 cartons. Find the probability that the total volume of juice in a pack of 6 cartons is less than 6 litres. Explain why this probability is less than your answer to part (i).

(6)

(iii) A new and more accurate machine is available, for which the volume of juice dispensed is Normally distributed but with smaller standard deviation 4 ml. By how much could the existing mean volume of juice dispensed into each carton be reduced without increasing the existing proportion of cartons with less than the nominal volume? Supposing that the additional cost of the more accurate machine is £200, and the cost of the apple juice is £1 per litre, how many cartons of juice would have to be filled by the more accurate machine in order to justify its greater cost?

(10)

4. (i) State the Normal approximation to the binomial distribution, indicating the conditions under which it is valid. In 50 firings, a surface-to-air missile (SAM) is successful in hitting its target in 30 cases. Obtain an approximate 95% confidence interval for the probability, p say, that a given missile hits its target.

(8)

(ii) SAMs are routinely fired at a target independently in pairs; if at least one missile hits the target, the target is destroyed. (It can be assumed that the two SAMs are successful in hitting the target independently of one another.) Find in terms of p the probability that a target is destroyed when a pair of missiles is fired at it, and provide a point estimate of this probability.

(3)

(iii) By suitably transforming the confidence interval for p in part (i), obtain a 95% confidence interval for the probability that the target is destroyed when a pair of missiles is fired.

(4)

(iv) Defence experts say that an airborne enemy target can be detected just before entering national airspace. The defence ministry has a policy that several pairs of SAMs (n pairs, say) should be fired whenever such a target is detected. Assuming that your estimate in part (ii) is accurate, find the smallest value of n which will reduce the probability of failing to destroy the target to below 0.0005.

(5)

Turn over

5. (i) The continuous random variable T has probability density function (pdf) $f(t)$ given by

$$f(t) = \lambda e^{-\lambda t}, \qquad t > 0; \quad \lambda > 0.$$

(a) Sketch the graph of $f(t)$.

(b) Show that the cumulative distribution function is given by

$$F(t) = 1 - e^{-\lambda t}, \qquad t > 0; \quad \lambda > 0.$$

(c) Deduce that

$$P(a < T \leq b) = e^{-\lambda a} - e^{-\lambda b}, \quad 0 < a < b.$$

(7)

(ii) The accounts manager for a building firm assumes that the time T taken to settle an invoice is a random variable with the above pdf $f(t)$ for some unknown value of λ. For a random sample of 100 invoices, he finds that 50 are settled within one week, 35 are settled during the second week and 15 are settled after 2 weeks. Explain clearly why the likelihood of these data may be written as

$$L(\lambda) = k\left(1 - e^{-\lambda}\right)^{50}\left(e^{-\lambda} - e^{-2\lambda}\right)^{35}\left(e^{-2\lambda}\right)^{15},$$

where k is a constant.

Hence show that

$$\log L(\lambda) = \log(k) + 85\log\left(1 - e^{-\lambda}\right) - 65\lambda .$$

Deduce that the maximum likelihood estimate of λ is approximately 0.836.

(8)

(iii) Assuming that $\lambda = 0.836$, calculate each of the expected numbers of invoices settled within the first week, settled during the second week, and settled after 2 weeks. Hence comment briefly on how well the model pdf $f(t)$ fits the data.

(5)

6. (i) The random variable X has probability density function $f(x)$ given by

$$f(x) = \frac{k}{x^{k+1}}, \qquad x \geq 1, \quad k > 0.$$

Sketch the graph of $f(x)$ and find the cumulative distribution function $F(x)$.

(5)

(ii) Find the median and the lower and upper quartiles of X and deduce the semi-interquartile range of X.

(5)

(iii) Assuming $k > 2$, find the expectation and variance of X. What is the probability that X exceeds its expectation?

(7)

(iv) In the country of Utopia, incomes in units of £10000 are distributed as is X with $k = 3$. Find (a) the median income, (b) the mean income, (c) the proportion of incomes greater than £100000.

(3)

7. The lengths X of offcuts of timber in a carpenter's workshop follow the continuous uniform distribution with probability density function (pdf)

$$f(x) = \frac{1}{\theta}, \qquad 0 \le x \le \theta,$$

where θ (> 0) is an unknown parameter.

(i) Find the mean and variance of X.

(5)

(ii) The carpenter takes a random sample of offcuts with lengths $X_1, X_2, ..., X_n$. Explain why

$$P\left(\text{length of longest offcut in sample } \le x\right) = \left(\frac{x}{\theta}\right)^n, \qquad 0 \le x \le \theta,$$

and deduce the pdf of the sample maximum, $X_{(n)}$ say. Show that

$$E\left(X_{(n)}\right) = \frac{n\theta}{n+1}$$

and

$$\text{Var}\left(X_{(n)}\right) = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Write down a multiple of $X_{(n)}$ which is an unbiased estimator of θ, and obtain its variance.

(11)

(iii) Show that $\frac{2}{n}\sum_{i=1}^{n} X_i$ is the method of moments estimator of θ, and obtain the variance of this estimator.

(4)

8. In a study of office efficiency, a firm has established benchmark times, $x_1, x_2, \ldots, x_{10}$, for the completion of 10 different routine office tasks. A newly recruited trainee is timed for his performance on each of these tasks, giving times $y_1, y_2, \ldots, y_{10}$. The times in minutes, (x_i, y_i), $i = 1, 2, \ldots, 10$, are tabulated below.

Task	A	B	C	D	E	F	G	H	I	J
Benchmark time x	5	5	10	10	10	15	15	20	20	40
Trainee's time y	8	12	16	16	21	18	20	25	31	53

Note: $\sum x_i = 150$, $\sum x_i^2 = 3200$, $\sum y_i = 220$, $\sum y_i^2 = 6280$, $\sum x_i y_i = 4440$.

(i) Plot a scatter diagram of these data and briefly comment on the suitability of simple linear regression analysis in this case.

(5)

(ii) Stating clearly your assumptions, use the method of least squares to fit a simple linear regression model to the data, and calculate (a) the residual mean square (the unbiased estimate of the variance σ^2 of the stochastic term in the regression model) and (b) the coefficient of determination, R^2.

(7)

(iii) Given that the variance of the slope estimate is $\dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}$, where $\bar{x}$ denotes the sample mean benchmark time, test at the 5% level of significance the null hypothesis that the slope of this regression is 1.

(4)

(iv) The office manager is dissatisfied with the regression relationship which you obtain. He believes that when $x = 0$ it should also be the case that $y = 0$. State the appropriate form of the linear regression model embodying this restriction, derive a formula for the estimate of its slope parameter, and estimate this parameter for the above data.

(4)