

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



HIGHER CERTIFICATE IN STATISTICS, 2005

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 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 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. In the UK National Lottery, players seek to guess six numbers, selected at random (without replacement) from the list 1, 2, 3, ..., 49.
- (i) Show that the total number of ways of choosing six numbers from 49 is 13,983,816. (2)
- (ii) Suppose that 10,000,000 people play the lottery, and assume that each player independently chooses, at random and without replacement, six numbers from the list 1, 2, 3, ..., 49. Let X denote the total number of players who match the six winning numbers. Write down the exact distribution of X , and a Poisson approximation to this distribution. Hence find, approximately, the probability
- (a) that $X = 0$,
- (b) that $X = 1$. (9)
- (iii) It is believed that many lottery players guess six numbers at random without replacement from the list 1, 2, 3, ..., 31. Calculate the total number of possible choices without replacement of six numbers from the list 1, 2, 3, ..., 31, and deduce the probability that the winning set of six numbers contains no number greater than 31. (4)
- (iv) Suppose now that 3,000,000 lottery players choose their six numbers, at random and without replacement, from the list 1, 2, 3, ..., 31, whilst 7,000,000 players choose their six numbers, at random and without replacement, from the list 1, 2, 3, ..., 49. Let Y denote the total number of players who match the six winning numbers. Write down the (approximate) distribution of Y
- (a) when all six winning numbers are in the list 1, 2, 3, ..., 31,
- (b) when at least one winning number is not in this list. (5)

2. I have three ways of travelling to work. If I cycle, my travelling time is distributed $N(27, 6.25)$, i.e. Normally with mean 27 minutes and standard deviation $\sqrt{6.25} = 2.5$ minutes. If I use the bus, my time walking from home to pick up the bus is distributed $N(7, 4)$, the bus journey time is distributed $N(13, 20)$ and the time to walk from the bus stop to work is distributed $N(5, 1)$, all three components of the journey time being independent. If I drive my car, the journey time is distributed $N(23, 36)$.
- (i) Find the distribution of my total journey time if I use the bus. (3)
- (ii) Which method of transport gives me the best chance of achieving a total journey time of 30 minutes or less, and what chance does it give me of doing so? (5)
- (iii) Which method of transport gives me the least chance of a journey time of 35 minutes or more, and what is then my chance of taking at least 35 minutes? (5)
- (iv) Taken over many journeys to work, the numbers of times I cycle, take the bus or drive the car are in the ratio 3 : 3 : 4. Given that my journey time yesterday was under 30 minutes, find the respective probabilities of the three modes of travel. (7)

3. The failures of a communications system occur in a Poisson process with rate parameter λ , so that the random variable X giving the number of failures in time t satisfies

$$P(X = x) = \exp(-\lambda t) \frac{(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \dots$$

- (i) By considering the probability that there is no failure in time t , or otherwise, show that the probability distribution of the time T to the next failure is given by

$$P(T > t) = \exp(-\lambda t), \quad t > 0,$$

and deduce the probability density function (pdf) of T .

(4)

- (ii) A bank of n identical but independent communications systems of the above type is set in operation at time 0, and T_i denotes the time to failure of the i th system, $i = 1, 2, \dots, n$. Obtain an expression for the probability that all n systems continue to function without failure for at least a time t . Deduce that the pdf of $T_{\min} = \min(T_1, \dots, T_n)$, the time to the first failure in the bank of n systems, is of exponential form, with a rate parameter given by a function of n and λ which should be stated.

(5)

- (iii) Show that the probability that all n systems in the bank have failed by time t is given by $(1 - e^{-\lambda t})^n$, and deduce the pdf of $T_{\max} = \max(T_1, \dots, T_n)$, the time to the last failure in the bank of n systems.

Given that $n = 10$ and $\lambda = 0.002$, find the times t_1 and t_2 such that

$$P(T_{\min} < t_1) = P(T_{\max} > t_2) = 0.05.$$

(11)

4. (i) The continuous random variable X is distributed with probability density function $f(x)$ given by

$$f(x) = \alpha(1-x)^{\alpha-1}, \quad 0 < x < 1, \quad \alpha > 0.$$

Find the cumulative distribution function of X , $F(x)$ say, and hence obtain the median of X . Also sketch the graphs of $f(x)$ and $F(x)$ for the case $\alpha = 3$.

(8)

- (ii) A random sample x_1, x_2, \dots, x_n is taken from this distribution with a view to estimating the unknown parameter α . Write down the likelihood function of these data, $L(x_1, x_2, \dots, x_n | \alpha)$ say, and show that the maximum likelihood estimate (MLE) of α is given by

$$\hat{\alpha} = \frac{-n}{\sum_{i=1}^n \log(1-x_i)}.$$

(4)

Also obtain $\frac{d^2 \log L(x_1, x_2, \dots, x_n | \alpha)}{d\alpha^2}$ as a function of α and n . Assuming that

$\hat{\alpha}$ is approximately Normally distributed with mean α and variance $-1/\left(\frac{d^2 \log L(x_1, x_2, \dots, x_n | \alpha)}{d\alpha^2}\right)$, deduce an approximate 90% confidence

interval for α . Evaluate this interval for the sample 0.12, 0.43, 0.07, 0.87, 0.29.

(8)

5. (i) A social scientist is conducting a survey of drug-taking among students. Her survey is based on face-to-face interviews with a random sample of n students from Wackford Squeers University, y of whom tell her that they take drugs. Assume that the true proportion of drug-takers in the population being sampled is p , that the responses are truthful, and that the number in the sample who take drugs follows the binomial distribution $B(n, p)$. Write down the likelihood function for these data and find the maximum likelihood estimator (MLE) of p , \hat{p} say. Also write down $\text{Var}(\hat{p})$ and hence obtain an estimate of the standard error of \hat{p} , $\text{SE}(\hat{p})$ say. Calculate the values of \hat{p} and $\text{SE}(\hat{p})$, given that $n = 100$ and $y = 20$.

(8)

- (ii) A statistician now advises her that, because of the sensitivity of the question of drug-taking, some students may not answer truthfully, so causing bias. He suggests using an anonymising device to encourage truthful answers. He provides a biased coin with faces labelled "takes drugs" and "does not take drugs", which show with respective probabilities 0.75 and 0.25 when the coin is tossed.

A second random sample, also of size n , is selected independently of the first. Each of the n students is asked to toss the coin, unseen by the interviewer, and to answer "yes" if he/she is in the group indicated by the coin and "no" otherwise.

Assuming truthful responses, show that the probability of a "yes", θ say, is given by

$$\theta = 0.25 + 0.5p .$$

Assuming further that the number of students answering "yes", Z say, is distributed $B(n, \theta)$ with observed value z , write down the MLE of θ , $\hat{\theta}$ say, and an estimate of its standard error. Use the relationship between θ and p to deduce the MLE of p , \tilde{p} say, and $\text{SE}(\tilde{p})$. Calculate the values of \tilde{p} and $\text{SE}(\tilde{p})$, given that $n = 100$ and $z = 45$.

(8)

- (iii) A journalist comparing the results of the two surveys says that the first survey is more reliable. Do you agree? Why, or why not?

(4)

6. The random variable X has the geometric probability mass function (pmf) given by $f(x)$, where

$$f(x) = (1-p)^x p, \quad x = 0, 1, 2, \dots, \quad 0 < p < 1.$$

- (i) Sketch the graph of $f(x)$ for the case $p = 1/3$, for $0 \leq x \leq 5$. (4)
- (ii) Show that the probability generating function of X is given by $G(s)$ where

$$G(s) = \frac{p}{1-(1-p)s}, \quad |s| < \frac{1}{1-p},$$

and hence or otherwise obtain the mean and variance of X . (6)

- (iii) For any non-negative integer x , show that $P(X \geq x) = (1-p)^x$, and deduce that for any non-negative integers l and m

$$P(X \geq l+m | X \geq l) = P(X \geq m).$$

Interpret this result. (4)

- (iv) The random variable Y has pmf $g(y)$, where

$$g(y) = (1-\theta)^y \theta, \quad y = 0, 1, 2, \dots, \quad 0 < \theta < 1.$$

X and Y are independent, and the random variable Z is defined as the minimum of X and Y , i.e. $Z = \min(X, Y)$. By noting that $P(Z \geq z) = P(X \geq z \text{ and } Y \geq z)$, find an expression for $P(Z \geq z)$, where z is any non-negative integer. By considering $P(Z \geq z) - P(Z \geq z+1)$, or otherwise, show that

$$P(Z = z) = [(1-p)(1-\theta)]^z (p + \theta - p\theta), \quad z = 0, 1, 2, \dots$$

Identify the form of this distribution and hence write down $E(Z)$ and $\text{Var}(Z)$. (6)

7. The breaking stresses in newtons per square metre of standard samples of pine, measured at varying angles x° to the grain of the wood, are tabulated below.

Breaking Stress y (N/m^2) $\times 10^{-10}$	0.987	1.064	1.337	1.912	2.740	5.771	11.494
Angle x°	0	15	30	45	60	75	90

- (i) Plot these data, use the summary information given below to calculate the product-moment coefficient of correlation $\text{corr}(x, y)$ between x and y , and comment briefly. (7)

- (ii) Hankinson's formula for y in terms of x is of the form

$$y = \left(\frac{1 - \sin^2 x}{a} + \frac{\sin^2 x}{b} \right)^{-1},$$

where a and b are the standardised breaking stresses parallel to and perpendicular to the grain respectively. Show how this relationship may be expressed in the linear form

$$Y = A + BX,$$

where $Y = 1/y$ and $X = \sin^2 x$ and A and B are functions of the unknown parameters a and b which you should find.

(3)

- (iii) Plot Y against X , comment on the suitability of this relationship for regression analysis, estimate A and B by least squares and deduce the corresponding estimates of a and b . Also compute $\text{corr}(X, Y)$ and compare this with the correlation computed in part (i).

(10)

Note: (A) State clearly any formulae assumed without proof.

(B) $\sum x = 315, \sum x^2 = 20475, \sum y = 25.305, \sum y^2 = 180.474, \sum xy = 1773.795,$
 $\sum \sin^2 x = 3.5, \sum \sin^4 x = 2.75, \sum \left(\frac{1}{y}\right) = 3.849, \sum \left(\frac{1}{y^2}\right) = 2.9136, \sum \frac{\sin^2 x}{y} = 1.03385.$

8. The joint probability mass function of X and Y is tabulated below.

		Values of Y		
		0	1	2
Values of X	-1	1/6	1/12	1/12
	0	1/12	1/6	1/12
	1	1/12	1/12	1/6

- (i) Obtain the marginal distributions of X and Y , and hence calculate $E(X)$, $E(Y)$, $\text{Var}(X)$ and $\text{Var}(Y)$. (5)
- (ii) Obtain the conditional distribution of Y for each possible value of X , and hence show that $E(Y | X = x)$ is a linear function of x . (5)
- (iii) Find $E(XY)$ and deduce $\text{Cov}(X, Y)$ and $\text{corr}(X, Y)$. Are X and Y independent? (5)
- (iv) Find the probability distribution of $Z = X^3 + (Y - 1)^3$. (5)