

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



HIGHER CERTIFICATE IN STATISTICS, 2007

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. When the digits 1, 2, 3, 4 are placed in ascending order (without repetition), each digit is said to occupy its "home" position. For example, the 3rd place is the home position of the digit 3.
- (i) How many possible orderings of the digits 1, 2, 3, 4 are there? (1)
- (ii) By enumeration, or otherwise,
- (a) show that there are 9 orderings in which no digit occupies its home position;
- (b) show that there are 8 orderings in which exactly one digit occupies its home position;
- (c) find the number of orderings in which exactly two digits each occupy their respective home positions;
- (d) find the number of orderings in which all four digits each occupy their respective home positions. (10)
- (iii) The digits 1, 2, 3, 4 are put in random order, and the random variable X denotes the number of digits occupying their respective home positions. Write down the probability distribution of X , and calculate $E(X)$ and $\text{Var}(X)$. (9)

2. State *Bayes' theorem*.

(4)

The gene for haemophilia (the inability of blood to clot) produces no symptom in females, but a female who carries the gene will pass it on to each of her children independently with probability $\frac{1}{2}$ in each case. Assume throughout this question that males with haemophilia (haemophiliacs) do not marry or have children.

(i) Mrs Smith is expecting her first child (a son), and it is known that Mrs Smith's maternal uncle (the only brother of Mrs Smith's mother) is a haemophiliac. Explain clearly why, in the light of this information, the probability that Mrs Smith's son will be a haemophiliac is $\frac{1}{8}$. What would be the probability if instead it is known that Mrs Smith's mother has four brothers, only one of whom is a haemophiliac?

(6)

(ii) Suppose throughout this part that Mrs Smith's mother has only one brother, who is a haemophiliac, and that Mrs Smith has four brothers, none of whom is a haemophiliac.

(a) Show that the probability that none of Mrs Smith's brothers is a haemophiliac, given that Mrs Smith's mother carries the gene, is $\frac{1}{16}$.

(b) Write down the probability that none of Mrs Smith's brothers is a haemophiliac, given that Mrs Smith's mother does not carry the gene.

(c) What is now the probability that Mrs Smith's son will be a haemophiliac?

(10)

3. The random variable X has the Pareto distribution with probability density function (pdf)

$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}}, \quad x > 0, \quad \alpha > 0,$$

and a random sample x_1, x_2, \dots, x_n is taken from this distribution.

- (i) Show that the maximum likelihood estimate of α is given by

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \log(1+x_i)}. \quad (5)$$

- (ii) Stating any results you assume without proof, show that $\hat{\alpha} \sim N(\alpha, \frac{\alpha^2}{n})$ approximately if the sample size is large. Hence derive an approximate 95% confidence interval for α in terms of $\hat{\alpha}$ and n . (5)

- (iii) A random sample x_1, x_2, \dots, x_{100} which is assumed to be from the pdf $f(x)$ gives $\sum_{i=1}^{100} \log(1+x_i) = 28.57$. Calculate $\hat{\alpha}$ and an approximate 95% confidence interval for α . (4)

- (iv) You are now given that, for $\alpha > 2$,

$$E(X) = \frac{1}{\alpha-1},$$

$$\text{Var}(X) = \frac{\alpha}{(\alpha-1)^2(\alpha-2)},$$

and that the sum and sum of squares of the sample values are 39.4 and 52.6 respectively. Evaluate the mean and variance of this distribution, using the value of $\hat{\alpha}$ found in part (iii). Also calculate the sample mean and variance, and comment briefly. (6)

4. The annual number, X say, of accidents on a certain stretch of road is assumed to follow a Poisson distribution with mean λ , so that

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots; \quad \lambda > 0.$$

For any accident on this stretch of road, the probability that it involves one or more fatalities (i.e. it is a fatal accident) is p , $0 < p < 1$, independently of all other accidents.

- (i) Given that a total of x accidents occurs in a given year, explain why the distribution of the number of fatal accidents in that year, $Y(x)$ say, is binomial, and write down the parameters of this distribution.

(4)

- (ii) Deduce that the unconditional probability that, in any given year, there are x accidents of which y are fatal, is given by

$$p(x, y) = \frac{e^{-\lambda}}{y!} [\lambda(1-p)]^{x-y} \frac{\lambda^y p^y}{(x-y)!},$$

for $x = 0, 1, 2, \dots$; $y = 0, 1, 2, \dots, x$; $\lambda > 0$.

(4)

- (iii) Hence show that the marginal distribution of the number of fatal accidents on this stretch of road in any given year is Poisson with mean λp .

(7)

- (iv) Given that $\lambda = 10$ and $p = 0.2$, find

- (a) the probability that in a given year there are no fatal accidents,
(b) the probability that, in a year of 6 accidents, at most one is fatal.

(5)

5. (i) The random variable X has the geometric probability mass function (pmf) given by

$$p(x; p_1) = (1 - p_1)^{x-1} p_1, \quad x = 1, 2, 3, \dots; \quad 0 < p_1 < 1.$$

Sketch the graph of $p(x; p_1)$ for the case $p_1 = 0.4$ and $1 \leq x \leq 5$.

(4)

- (ii) For any non-negative integer x , show that $P(X > x) = (1 - p_1)^x$.

(4)

- (iii) The random variable Y is distributed independently of X with pmf $p(x; p_2)$, where $0 < p_2 < 1$. By writing

$$P(X > Y) = \sum_{y=1}^{\infty} P(Y = y)P(X > Y | Y = y),$$

or otherwise, show that

$$P(X > Y) = \frac{p_2(1 - p_1)}{p_1 + p_2 - p_1 p_2},$$

and hence write down $P(X < Y)$.

(6)

- (iv) Show that

$$P(X = Y) = \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2},$$

and deduce a condition on p_1 and p_2 such that $P(X = Y) = \frac{1}{2}$.

If $P(X = Y) = \frac{1}{2}$ and also $P(X < Y) = P(X > Y)$, solve for p_1 and p_2 .

(6)

6. (i) A passenger aircraft carries 100 passengers. Assume that the weight in kg, W say, of a randomly chosen passenger is Normally distributed with mean 65 and standard deviation 6, i.e. $W \sim N(65, 36)$. Assume also that a randomly chosen passenger has hold luggage weighing H kg, where $H \sim N(20, 4)$, and cabin luggage weighing C kg, where $C \sim N(6, 9)$, W , H and C being independent within and between passengers.
- (a) Write down the distribution of the total weight in kg, T say, of a randomly chosen passenger and his or her hold and cabin luggage, and find $P(T > 110)$. (4)
- (b) A safety requirement is that the total weight of all 100 passengers and their luggage should not exceed 9300 kg. Find the probability that this requirement is not met. (3)
- (c) The check-in scales record the weight of hold luggage to the nearest kg. Find the probability that a randomly chosen passenger's hold luggage is recorded as weighing 24 kg or more. (3)
- (ii) Now suppose that, for each passenger, the random variables H and C are not independent but instead the correlation between them is -0.5 , all other assumptions being as before. What is now the probability that the safety requirement is not met? (6)
- (iii) Comment critically on the assumptions listed in part (i) of this question. (4)

7. (a) The random variable X has the binomial distribution with probability mass function (pmf)

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n; \quad 0 < p < 1.$$

Show that, for $x = 1, 2, \dots, n$,

$$f_X(x) = \frac{p(n+1-x)}{(1-p)x} f_X(x-1),$$

and deduce that the mode m of X is the largest integer satisfying

$$m \leq (n+1)p.$$

By considering the case $\frac{p(n+1-x)}{(1-p)x} = 1$, write down an example of a binomial distribution which has two consecutive equal maximal probabilities, and show that, for any such distribution, the mean np lies between the two modes.

(10)

- (b) In a game of target practice, a player has two guns, A and B , which must be fired alternately. His probability of hitting the target is 0.8 with A and 0.4 with B . He is allowed three shots at the target, which must be fired *either* in the order $A B A$ or in the order $B A B$. Assume that events relating to different shots are independent.

- (i) Which sequence should he choose to maximise the probability of hitting the target in two consecutive shots?
- (ii) Find the mean and variance of the number of times the player hits the target when firing in the order $A B A$; and when firing in the order $B A B$.
- (iii) Comment on your results.

(10)

8. A uniform thin wire is fixed at its top end. Varying loads (L , in hundreds of grams) are hung from the bottom end of the wire, and the extension (E , in millimetres) of the wire for each load is measured. The results are as shown in the following table.

L	1	1	2	2	3	3	4	4	5	5	$\Sigma L = 30$
E	2	3	4	2	8	10	13	13	11	14	$\Sigma E = 80$

$$\Sigma L^2 = 110 \quad \Sigma E^2 = 852 \quad \Sigma LE = 300$$

- (i) It is proposed to analyse these data using simple linear regression analysis. State which of L and E should be taken as the independent variable and which as the dependent variable, and why. Construct a scatterplot of the data and comment on their suitability for linear regression analysis.

(8)

- (ii) (a) Fit a straight line to the data by the method of least squares. Calculate the residual mean square, s^2 say, for this regression, and give a point estimate of the expected extension when the load is 350 grams.

Note: you should state clearly any formulae you use, but you do not need to prove them.

(8)

- (b) You are given that, when a linear regression of y on x is fitted to data $(x_1, y_1), \dots, (x_n, y_n)$, the standard error of the intercept parameter is

$$s \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}},$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Theoretical physics suggests that the extension of

the wire should be proportional to the load put on it, which implies that the intercept parameter in the linear regression fitted in (i) should be zero. Test this hypothesis against a two-sided alternative, stating any necessary assumptions.

(4)