EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2003

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* ${}^{n}C_{r}$.

This examination paper consists of 10 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. A bank sort code consists of six digits.

(i) How many different sort codes may be formed from

(a) six digits chosen from 0, 1, 2, …, 9 allowing repetition,

(b) six different digits chosen from 0, 1, 2, …, 9,

(c) six different digits chosen from 0, 1, 2, …, 9, restricted to be in ascending order,

(d) three different digits chosen from 0, 1, 2, …, 9, each digit being used twice?

(8)

(ii) A sort code is palindromic if reversing it gives the same code. (For example, both 123321 and 946649 are palindromic sort codes, but 123312 is not.) It follows that a palindromic sort code can involve at most three different digits. How many palindromic sort codes can be formed using some or all of the digits 0, 1, 2

(a) if each digit is used twice,

(b) if one digit is used four times,

(c) if one digit is used six times?

(6)

(iii) How many palindromic sort codes can be formed altogether, using digits chosen from 0, 1, 2, …, 9?

(6)

2. The events A, B and C have respective probabilities $\frac{2}{3}$, $\frac{1}{2}$ and $\frac{1}{4}$, and $\overline{A}$, $\overline{B}$ and $\overline{C}$ are, respectively, the complements of A, B and C.

(i) Given that A, B and C are mutually independent, find

(a) $P\left(A \cap \overline{B} \cap \overline{C}\right)$,

(b) $P\left(A \cap \overline{C} \middle| A \cap \overline{B}\right)$.

(8)

(ii) Now let A and B be independent, A and C be independent and B and C be independent, so that A, B and C are pairwise independent, and let $P\left(A \cap B \cap C\right) = x$. Find in terms of x

(a) $P\left(A \cap \overline{B} \cap \overline{C}\right)$,

(b) $P\left(A \cap \overline{C} \middle| A \cap \overline{B}\right)$,

(c) $P\left(A \cup B \cup C\right)$.

Also find the maximum and minimum possible values of x.

(12)

3. Hand-crafted widgets are made in two types, A and B. All widgets require an initial process which takes a time X which is distributed as N(10, 12), i.e. Normally with mean 10 minutes and variance 12 minutes^2. Type A widgets take a further random time Y_A to complete, distributed N(15, 16), and type B widgets take a further time Y_B, distributed N(12, 9), where X, Y_A and Y_B are mutually independent.

(i) State the distribution of the total time to make

(a) a type A widget,

(b) a type B widget.

(5)

(ii) Two particular widgets take the same time in the initial process. If one is of type A and the other of type B, find the probability that the type A widget is completed first.

(5)

(iii) Find the probability that a randomly chosen type A widget has a shorter total completion time than a randomly chosen type B widget.

(5)

(iv) Widgets of each type are packed in boxes of 16. Find the probability that the sample mean completion time for a box of type A widgets is less than that for a box of type B widgets.

(5)

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

$$f(x) = kx^2(1-x), \qquad 0 \le x \le 1.$$

(i) Show that $k = 12$.

(2)

(ii) Show that the mode of X is at $x = \frac{2}{3}$ and draw a graph of $f(x)$.

(5)

(iii) Find the mean and variance of X.

(5)

(iv) Find the cumulative distribution function of X and obtain the probability that X lies within one standard deviation of its mean.

(8)

5. (i) The random variable X follows a Poisson distribution with probability mass function

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

Sketch $f(x)$ for the cases $\lambda = 0.5$ and $\lambda = 2$, and state the expectation and variance of X.

(6)

(ii) Given a random sample $x_1, x_2, \ldots, x_n$ from this distribution, obtain the maximum likelihood estimator of λ, $\hat{\lambda}_{ML}$ say. State a suitable approximation for the distribution of $\hat{\lambda}_{ML}$ assuming that the sample size n is large, and use this approximation to deduce an approximate 95% confidence interval for λ.

(7)

(iii) A random sample of 400 observations yields $\Sigma x_i = 2500$. Calculate an approximate 95% confidence interval for λ. Given further that $\Sigma x_i^2 = 25600$, calculate the sample variance of the given sample. Recalculate the confidence interval for λ using the central limit theorem but without assuming that the data are Poisson distributed. Compare this interval with that found in part (ii) and comment briefly.

(7)

6. The random variable Y follows a binomial distribution with probability mass function given by

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, ..., n; \quad 0 < p < 1.$$

Write down the mean and variance of Y.

(2)

(i) An intelligence test consists of 48 multiple-choice questions. For each question, four possible answers are presented but only one is correct. If a student answers all the questions independently by random guesswork, what will be the distribution of the number of questions he gets right?

(2)

(ii) Assume that, for any question, if a student knows the answer he writes it down correctly, and otherwise he guesses at random. If he knows the answer to 36 questions, find

(a) the mean and variance of the number of questions he gets right,

(b) the distribution of the number of questions he gets wrong,

(c) the probability that he gets more than two questions right by chance.

(8)

(iii) The test is given separately to students A, B and C in a tutor-group who know the answers to 27, 28 and 30 questions respectively. Find the mean and variance of the average number of questions they get right. Given that an unnamed test paper (which is *a priori* equally likely to be from any one of these students) has 29 questions right, find the respective probabilities that this paper was written by A, B or C.

(8)

7. (i) The random variable X denotes the number of failures preceding the first success in a series of independent Bernoulli trials, in each of which the probability of success is p. Show that the probability mass function of X is

$$P(X = x) = p(1-p)^x, \qquad x = 0, 1, 2, \ldots .$$

and draw a graph of this function for the case $p = 0.4$.

(6)

(ii) Show that the probability generating function of X is given by

$$G_X(s) = \frac{p}{1-(1-p)s} .$$

[for a suitable range of values of s].

Hence or otherwise obtain the mean and variance of X.

(8)

(iii) The random variable Y is defined as the number of trials up to and including the first success in the series of Bernoulli trials referred to in part (i). Express Y in terms of X, write down the probability mass function and the probability generating function of Y and state the mean and variance of Y.

(6)

8. (i) Given a random sample of paired data (x_1, y_1), (x_2, y_2), …, (x_n, y_n), write down the formula for the sample product-moment correlation coefficient, r say, and explain the meaning of this quantity. Also draw scatter diagrams to illustrate

(a) strong positive correlation,

(b) strong negative correlation,

(c) independent data,

(d) data that are uncorrelated but not independent.

(8)

(ii) (a) The following two pages **(pages 9 and 10)** show edited Minitab analysis of the relationship between cholesterol level and age for a sample of 9 males. Name the statistical model which is being fitted, identify the independent and dependent variables and state the usual assumptions made about the data.

(b) Use the output to calculate the correlation between "chol" and "age", and the correlation between "newchol" and "newage", and give a reason for the difference between these two correlations.

(c) Use the output to assess whether the constant term in the model is statistically significant. What would omission of the constant term imply for a new-born infant?

(d) Which of the ("chol", "age") and ("newchol", "newage") analyses do you prefer as a summary of the data, and why?

(12)

Edited Minitab output for this question follows on pages 9 and 10

Turn over

```
Cholesterol (mg/ml)(c1) 2.20  3.55  2.40  2.55  2.85  2.95  3.05  3.10  3.45
Age (years) (c2)        26    27    28    29    31    33    35    38    41
MTB > gstd
MTB > name c1 'chol' c2 'age'
MTB > plot c1 c2
          -       *
          -                                                       *
      3.50+
          -
 chol     -
          -
          -                                     *         *
      3.00+                           *
          -                     *
          -
          -
          -               *
      2.50+
          -           *
          -
          -    *
          -
            ------+---------+---------+---------+---------+---------+age
               27.0      30.0      33.0      36.0      39.0      42.0
MTB > regress c1 1 c2
The regression equation is    chol = 1.30 + 0.0500 age

Predictor       Coef       StDev          T        P
Constant      1.3000      0.8851       1.47    0.185
age          0.05000     0.02734       1.83    0.110

S = 0.4000      R-Sq = 32.3%     R-Sq(adj) = 22.7%

Analysis of Variance
Source       DF          SS          MS         F        P
Regression    1      0.5350      0.5350      3.34    0.110
Error         7      1.1200      0.1600
Total         8      1.6550

Unusual Observations
Obs       age       chol        Fit  StDev Fit   Residual    St Resid
  2      27.0      3.550      2.650      0.191      0.900       2.56R
MTB > copy c1 c2 c3 c4; SUBC> omit row 2.
MTB > name c3 'newchol' c4 'newage'
MTB > plot c3 c4
      3.50+                                                       *
          -
 newchol  -
          -
          -                                     *         *
      3.00+                           *
          -                     *
          -
          -
          -               *
      2.50+
          -           *
          -
          -    *
          -
      2.00+
            ------+---------+---------+---------+---------+---------+newage
               27.0      30.0      33.0      36.0      39.0      42.0
```


Turn over

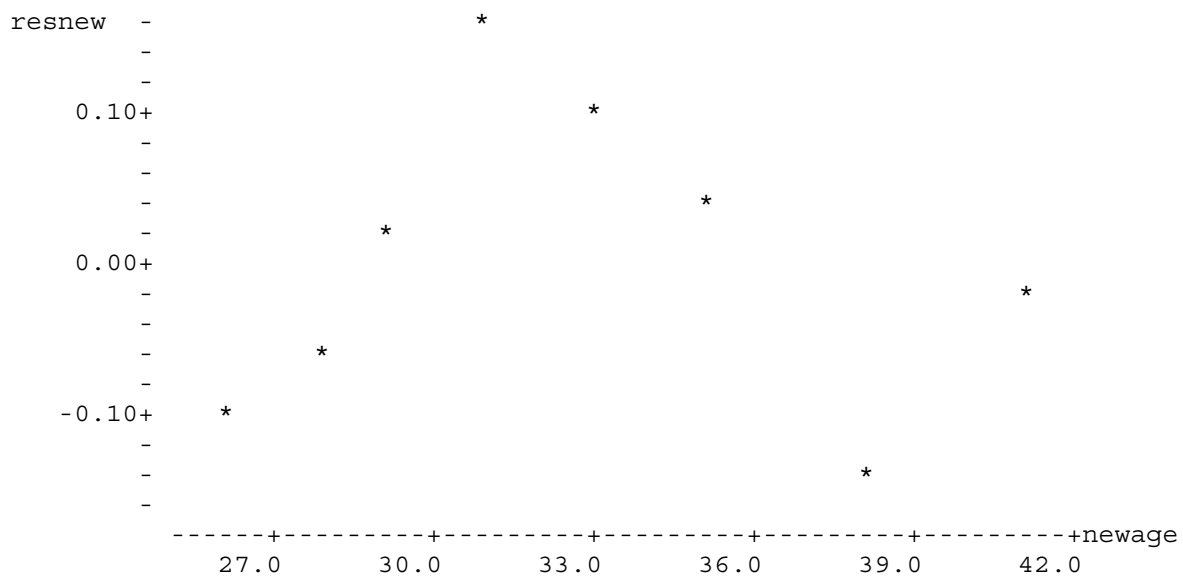
```
MTB > regress c3 1 c4;  SUBC> resi c5.
The regression equation is    newchol = 0.299 + 0.0772 newage

Predictor       Coef       StDev          T        P
Constant      0.2989      0.2629       1.14    0.299
newage      0.077236    0.007971       9.69    0.000

S = 0.1087      R-Sq = 94.0%     R-Sq(adj) = 93.0%

Analysis of Variance
Source       DF          SS          MS         F        P
Regression    1      1.1088      1.1088     93.88    0.000
Error         6      0.0709      0.0118
Total         7      1.1797

MTB > name c5 'resnew'
MTB > plot c5 c4
```




```
MTB > nsco c5 c6
MTB > plot c5 c6; SUBC> title 'Normal probability plot'.
```

Normal probability plot