EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2000

CERTIFICATE IN OFFICIAL STATISTICS, 2000

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.*

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

This examination paper consists of 12 printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

1. An ecological study is to be made of birds of an endangered species. When a bird is captured for the first time, three small, light, coloured bands are put on each leg, in upper, middle and lower positions, and the bird is then released. Each band may be red, yellow, blue or white.

   (i) If, regardless of the colours of other bands, each band position on either leg may have a band of any of the four colours, show that the total possible number of different colour combinations is 4096.

   (5)

   (ii) How many different colour combinations are possible if adjacent colour bands on each leg are restricted to be of different colours?

   (5)

   (iii) How many different colour combinations are possible if on each leg the three colours must all be different?

   (5)

   (iv) The colour sequence (upper, middle, lower) on the right leg is now restricted to be different from that on the left leg. Subject to this overriding condition, calculate the numbers of different possible colour combinations in the cases (i), (ii) and (iii) respectively.

   (5)

2. (a) Suppose that a lawyer has been guilty of financial irregularities in 5 of the 50 client accounts that she controls. An auditor randomly samples 10 accounts to check in detail. Find the probability that the auditor checks

(i) none,
(ii) two or more of the irregular accounts.

(5)

(b) In tennis, when the score reaches deuce ("40 all") the game is won by the first player to lead by two consecutive points. Suppose that the outcomes of all points are independent, and that the server is twice as likely to win a point as the receiver. Show that, once a game has reached deuce, the probability that the server wins the game is $\frac{4}{5}$.

(4)

Writing N for the number of points played from when the game first reaches deuce until it ends (i.e. is won by server or receiver), show that

$$P(N=n)=\frac{5}{4}\left(\frac{2}{3}\right)^{n}$$

and state the range of possible values for N.

(4)

(c) In the large city of Olchester, 30% of electors are Conservatives, 40% are Labour supporters, 20% are Liberal Democrats and 10% have no affiliation. Political affiliation is independent of sex, so these percentages apply to males and females equally. Records show that in a particular election 80% of the Conservatives voted, as did 60% of Labour supporters and 90% of Liberal Democrats, whilst those with no affiliation did not vote. If an elector is chosen at random and it is found that he did not vote in the election, find the probability that he is a Labour supporter.

(5)

A second elector is chosen at random and it is found that she also did not vote; what is the probability that both people are Labour supporters?

(2)

3. (a) For $i = 1, 2, \ldots, n$, let X_i be Normally distributed with mean μ_i and variance σ_i^2, and let $X_1, X_2, \ldots, X_n$ be independent. If $a_1, a_2, \ldots, a_n$ are any non-zero constants, state the distribution of $Y = \sum_{i=1}^{n} a_i X_i$. Deduce the distribution of $X_1 + X_2$ and $X_1 - X_2$.

(4)

(b) An office manager takes, on average, 35 minutes to get to work; his travelling time is Normally distributed with standard deviation 4 minutes. His secretary takes, on average, 33 minutes to get to work; her travelling time is Normally distributed with standard deviation 3 minutes. Stating any further assumptions you make, find

(i) the probability that the manager reaches work before his secretary, supposing that they leave their homes at the same time,

(3)

(ii) how much earlier the secretary must leave her home to be 90% certain of getting to work before her manager,

(3)

(iii) the probability that both get to work within 30 minutes of the times at which they set out.

(3)

(c) In an office there are 20 computers, each of which has, independently of the rest, a Poisson incidence of breakdowns at the rate of 0.02 per week. Breakdowns are invariably repaired with negligible delay. Find the probabilities that

(i) a period of 4 weeks passes with no breakdowns,

(2)

(ii) in a period of 4 weeks there is at least one breakdown each week.

(2)

Use a suitable approximation to calculate the probability of more than 26 breakdowns over a 52 week period.

(3)

4. Two players, *A* and *B*, each simultaneously and independently roll a fair die. Let *X* and *Y* be random variables denoting the respective scores of *A* and *B* on any given roll, so that

$$P\left(X = x, Y = y\right) = \begin{cases} \dfrac{1}{36} & x, y = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

(i) Show that

$$P\left(X = Y\right) = \frac{1}{6}$$

and that

$$P\left(X > Y\right) = \frac{5}{12} \quad .$$

(8)

(ii) Let *Z* be a random variable denoting the number of times *A* and *B* each have to roll their dice for one or both to score a six. Explain why

$$P\left(Z = z\right) = \begin{cases} \dfrac{11}{36}\left(\dfrac{25}{36}\right)^{z-1} & z = 1, 2, 3, ... \\ 0 & \text{otherwise} \end{cases}$$

(4)

(iii) Making use of this result, find

(a) $P\left(Z \le 4\right)$,

(b) $E\left(Z\right)$,

giving your answers to 3 significant figures.

(8)

5. It is desired to carry out a blood test on a large number (N) of persons, to check for the presence or absence of a rare characteristic. It is assumed that the probability p of a positive result is the same for all persons and is independent of the results of tests on other persons.

To reduce the work of testing, the blood samples of N persons are pooled into m groups of size k, where $N = mk$. The k samples in each group are first tested together. If the group test is negative, no further test is necessary for the persons in that group. If the group test is positive each person is tested individually and so in all (k + 1) tests are required for the group of k persons.

(i) Find the probability that the test for any particular pooled sample of k persons is positive.

(4)

(ii) Let S_k denote the total number of tests required for the N people when initially tested in groups of k. Explain why S_k can be written in terms of a binomial random variable X, as

$$S_k = m + kX, \quad \text{where } X \sim B\left(m, 1-(1-p)^k\right) .$$

(3)

(iii) Hence show that

$$E\left(S_k\right) = N\left[\frac{1}{k} + 1 - (1-p)^k\right] ,$$

$$\mathrm{Var}\left(S_k\right) = Nk(1-p)^k\left[1-(1-p)^k\right] .$$

(3)

(iv) Assuming that p = 0.01, show by calculation that $E\left(S_{10}\right) > E\left(S_{11}\right)$ and that $E\left(S_{11}\right) < E\left(S_{12}\right)$.

(3)

(v) Find the value of k for which $E\left(S_k\right)$ is minimised when p = 0.05.
[Note: the optimum value of k lies between 4 and 7 inclusive.]

(3)

(vi) Compare your results for parts (iv) and (v) with those based on k = 1 (i.e. no pooling), and comment briefly.

(4)

6. The total claim amount *X* made in one year on a portfolio of insurance policies has probability density function

$$f(x) = \begin{cases} \lambda^2 x e^{-\lambda x} & x \geq 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

(i) Show that

(a) $E(X) = \dfrac{2}{\lambda}$,

(b) $\text{Var}(X) = \dfrac{2}{\lambda^2}$,

(c) $P(X > x) = e^{-\lambda x}(1 + \lambda x)$.

(5)

(ii) If *X* is measured in units of £1000, λ may be assumed to take the value 0.01. The company has a total sum (policyholders' premiums + reserves) of £500,000 available to meet the year's claims. Show that the probability that the company is ruined (i.e. $P(X > 500)$) is 0.040 (to 3 decimal places).

(5)

(iii) A trainee actuary mistakenly assumes the distribution of total claim amount to be Normal with the same mean and variance as *X* (taking $\lambda = 0.01$). On this assumption, find the probability of ruin, given that £500,000 is available to meet the year's claims.

(5)

(iv) Making this mistaken assumption of Normality, the trainee calculates that £450,000 is the sum to be set aside to meet the year's claims with a probability of ruin of less than 0.04. What is the true probability of ruin, if only £450,000 is available to meet the year's claims?

(5)

7. The random variable X follows the discrete uniform distribution on the integers 1, 2, ..., k so that the probability mass function of X is given by

$$p(x) = \begin{cases} \dfrac{1}{k} & x = 1, 2, ..., k \\ 0 & \text{otherwise} \end{cases}$$

(i) Show that for a general positive integer k

$$E(X) = \frac{k+1}{2} \ .$$

(5)

(ii) A random sample of size four, X_1, X_2, X_3, X_4, is taken from this distribution, yielding values x_1, x_2, x_3, x_4, from which it is intended to estimate the parameter k.

(a) Show that the method of moments estimator of k is given by $\hat{k}_1 = 2\overline{X} - 1$, where $\overline{X}$ denotes the sample mean.

(5)

(b) Explain clearly why the maximum likelihood estimator of k is given by $\hat{k}_2 = X_{(4)}$, the sample maximum.

(5)

(c) Calculate $\hat{k}_1$ and $\hat{k}_2$ for the sample (1, 10, 3, 2) and comment on your results.

(5)

Turn over

8. The accompanying edited Minitab output shows some regression analysis of the progress in sales in 100s (*y*) of hi-tec widgets over time in months (*x*). Use the output to answer the following questions.

(i) In the output for Regression 1, the *p*-values for the *t* tests for the slope and intercept parameters are missing. Use available information to test the fitted parameters for statistical significance. Also use the output to calculate corr(*x*, *y*). Comment critically on the adequacy of this regression as a model for the data, having regard to Plots 1A and 1B.

(5)

(ii) Contrast the significance of the *x* term in Regression 2 with the result of your test for the *x* term in Regression 1. Interpret the statement 'R-Sq = 98.1%' in terms of the Analysis of Variance output for Regression 2. What standard assumption(s) about the distribution of the error term may be called into question by Plot 2? What other reason is there for not accepting Regression 2 as an adequate model for the data?

(5)

(iii) Critically compare Regressions 1 and 3 for their success in fitting the data. Interpret the statement 'R-Sq = 99.1%' in Regression 3 in terms of the Analysis of Variance output, and explain why the total sum of squares in Regression 3 (0.28760) differs from the total sum of squares in Regressions 1 and 2 (186120).

(5)

(iv) Use each of the three regressions to give point estimates of sales at 10 months after launching the product. State with reasons which of the three estimates you think is the best.

(5)

Two pages of Minitab output follow

```
MTB > Print ’y’ 175  195  225  275  333  383  423  483  643 # sales of widgets (100s)
MTB > Print ’x’   1    2    3    4    5    6    7    8    9 # time from product launch (months)
MTB > Plot ’y’ ’x’.
         -
      640+                                                         *
         -
 y       -
         -
         -
      480+                                                  *                    PLOT 1A
         -                                                                       -------
         -                                            *
         -                                     *
         -
      320+                              *
         -                        *
         -
         -                 *
         -          *
      160+    *
           ------+---------+---------+---------+---------+---------+x
               1.5       3.0       4.5       6.0       7.5       9.0

MTB >  #################   REGRESSION  1   #################
MTB > Regress ’y’ 1 ’x’; SUBC>  Residuals C4.  # C4 stores residuals from the regression of y on x
The regression equation is y = 78.3 + 54.0 x
Predictor        Coef       StDev          T        P
Constant        78.33       29.01       2.70
x              54.000       5.155      10.48
S = 39.93       R-Sq = 94.0%     R-Sq(adj) = 93.1%
Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1      174960      174960    109.74    0.000
Residual Error     7       11160        1594
Total              8      186120
MTB >  Let C5 = 78.33 + 54*’x’     # C5 gives the fitted values from regression 1
MTB > Name  C3 ’x-sq’  C4 ’slr-res’ C5 ’slr-fit’
MTB > Plot ’slr-res’ ’slr-fit’.
       80+                                                        *
         -
 slr-res -
         -
         -
       40+  *                                                                    PLOT 1B
         -                                                                       -------
         -
         -
         -        *
        0+
         -
         -               *      *      *     *
         -                                                 *
         -                                          *
      -40+
         -
           ----+---------+---------+---------+---------+---------+--slr-fit
             160       240       320       400       480       560

MTB > Let C3 = ’x’*’x’       # C3 is time  squared
MTB >  #################   REGRESSION  2   #################
MTB > Regress ’y’ 2  ’x’ ’x-sq’; SUBC>   Residuals C6.
The regression equation is
y = 170 + 4.0 x + 5.00 x-sq
Predictor        Coef       StDev          T        P
Constant       170.00       30.56       5.56    0.001
x                4.00       14.03       0.29    0.785
x-sq            5.000       1.368       3.65    0.011
S = 24.01       R-Sq = 98.1%     R-Sq(adj) = 97.5%
Analysis of Variance
Source            DF          SS          MS         F        P
Regression         2      182660       91330    158.38    0.000
Residual Error     6        3460         577
Total              8      186120
Source       DF      Seq SS
x             1      174960
x-sq          1        7700
MTB > Let C7=170 + 4*’x’ + 5*’x-sq’    #  C7 stores fitted values from regression 2
MTB > Name C6 ’quadrres’ C7 ’quadrfit’
MTB > Plot ’quadrres’ ’quadrfit’.
         -                                                       *
       25+
         -                  *
 quadrres-                                                                       PLOT 2
         -            *             *                                            ------
         -
        0+       *
         - *  *
         -
         -
         -                                  *
      -25+
         -
         -
         -                                            *
         -
           --------+---------+---------+---------+---------+--------quadrfit
                 240       320       400       480       560

MTB > Let C8 = logten(’y’)      # C8 contains log (y) to the base 10
MTB > Name C8 ’log10(y)’
MTB > Plot ’log10(y)’ ’x’.
         -
     2.80+                                                         *
         -
 log10(y)-
         -                                                  *
         -                                            *
     2.60+                                     *                                 PLOT 3A
         -                                                                       -------
         -                              *
         -
         -                        *
     2.40+
         -                 *
         -
         -          *
         -    *
     2.20+
           ------+---------+---------+---------+---------+---------+x
               1.5       3.0       4.5       6.0       7.5       9.0

MTB >  #################   REGRESSION  3   #################
MTB > Regress ’log10(y)’ 1  ’x’; SUBC>    Residuals C9.   # C9 stores residuals from regression 3
The regression equation is
log10(y) = 2.16 + 0.0689 x
Predictor        Coef       StDev          T        P
Constant      2.16086     0.01422     151.93    0.000
x            0.068910    0.002527      27.26    0.000
S = 0.01958     R-Sq = 99.1%     R-Sq(adj) = 98.9%
Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1     0.28492     0.28492    743.35    0.000
Residual Error     7     0.00268     0.00038
Total              8     0.28760
MTB > Let C10=2.16086 + 0.06891*’x’   # C10 stores fitted values from regression 3
MTB > Name C9 ’logy-res’  C10 ’logy-fit’
MTB > Plot ’logy-res’ ’logy-fit’.
 logy-res-
         -                                                     *
         -
    0.020+
         -                              *
         -       *                                                               PLOT 3B
         -                                    *                                  -------
         -                        *
    0.000+
         -
         -             *
         -
         -                  *                      *
   -0.020+
         -
         -                                               *
         -
           +---------+---------+---------+---------+---------+------logy-fit
        2.16      2.28      2.40      2.52      2.64      2.76
```