EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2001

CERTIFICATE IN OFFICIAL STATISTICS, 2001

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 14 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. The events A, B and C are such that

A is independent of B,

A is independent of C,

and

$$P(A) = \tfrac{2}{3} \ ,$$
$$P(B) = \tfrac{1}{2} \ ,$$
$$P(C) = \tfrac{3}{5} \ ,$$
$$P(A \cap B \cap C) = \tfrac{1}{4} \ ,$$
$$P(\overline{A} \cap \overline{B} \cap C) = \tfrac{1}{10} \ .$$

Find

(i) $P(\overline{A} \cap \overline{B})$,

(3)

(ii) $P(\overline{A} \cap \overline{B} \cap \overline{C})$,

(4)

(iii) $P(B \cap C)$,

(4)

(iv) $P(B \mid C)$,

(3)

(v) $P(A \mid B \cap C)$,

(3)

(vi) $P(A \cap B \mid A \cap C)$.

(3)

2. (a) Three married couples at a dinner party sit down at random at a circular table set for six people. Find the probability that

(i) each man sits between two women,

(4)

(ii) all three men sit together,

(4)

(iii) exactly two men sit together.

(4)

(b) A disease, which can only be diagnosed with certainty after death, exists in a proportion p_0 of the population. A clinical test is known such that

P(test is positive given disease is present) $= p_1$

and

P(test is negative given disease is absent) $= p_2$.

(i) Find, in terms of p_0, p_1 and p_2, the probability that a randomly chosen individual who tests positive actually has the disease.

(4)

(ii) Calculate the answer to (i) in the case $p_0 = 0.005$, $p_1 = 0.95$, $p_2 = 0.95$, and comment on your result.

(4)

3. A manufacturer produces components of two quality grades:

Standard quality, with lifetimes distributed as N(2000, 90000), i.e. Normally with mean 2000 hours and standard deviation $\sqrt{(90000)} = 300$ hours.

High quality, with lifetimes distributed as N(2500, 15625).

(a) (i) Find the probability that a randomly chosen standard component lasts at least 2300 hours, and find the corresponding probability for a high quality component.

(4)

(ii) Find the probability that a randomly chosen standard component lasts longer than a randomly chosen high quality component.

(4)

(b) Due to a machine malfunction, a large batch of components is produced, of which 60% are standard and 40% high quality; however, these components are unlabelled and indistinguishable in appearance. A single component is chosen at random from this batch.

(i) Find the expectation of its lifetime and the probability that it lasts at least 2600 hours.

(6)

(ii) Given that a component from this batch lasts at least 2600 hours, what is the probability that it is of the high quality type?

(3)

(iii) Given that the standard deviation of the lifetime of a component taken at random from this batch is 346.8 hours, find the approximate probability that the mean lifetime of 100 such components exceeds 2300 hours.

(3)

4. The random variable X follows the binomial B(n, p) distribution with probability mass function

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad , \quad x = 0, 1, ..., n \quad , \quad 0 < p < 1 \quad ,$$

where $q = 1 - p$. Show that $E(X) = np$ and $\text{Var}(X) = npq$.

(5)

A mathematics class in a school is divided into set A with 12 students and set B with 25 students. Both groups are given a test consisting of 16 short questions. For any student in set A, the score (that is, the number of correct answers) is distributed as B(16, 0.75); for any student in set B, the score is distributed as B(16, 0.5). All students answer independently.

(i) Find the probability that

(a) a given set A student gets all 16 questions right,

(3)

(b) at least one student in set A gets all 16 questions right.

(3)

(ii) Use an appropriate approximation to find the probability that a given set B student scores more than a given set A student.

(5)

(iii) Let $\bar{X}$ and $\bar{Y}$ denote the mean scores of students in set A and set B respectively. Write down $E(\bar{X})$ and $E(\bar{Y})$, and show that $\text{Var}(\bar{X}) = 1/4$ and $\text{Var}(\bar{Y}) = 4/25$.

(4)

5. In an experiment, the number of events happening in any unit time interval is a Poisson random variable X with probability mass function

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

Show that

$$p(x+1) = \frac{\lambda}{x+1}p(x)\ , \quad x = 0, 1, 2, \ldots\ .$$

(1)

Draw graphs of $p(x)$ for the cases $\lambda = ½$ and $\lambda = 2$.

(4)

Obtain the moment generating function (mgf) of X and hence show that $E(X) = \mathrm{Var}(X) = \lambda$. Show also that $E[(X - \lambda)^3] = \lambda$.

(8)

In successive unit time intervals, the numbers of events $X_1, X_2, \ldots, X_n$ are independent and each has the same distribution as X. Obtain the mgf of $Y = X_1 + X_2 + \ldots + X_n$, deduce the form of the distribution of Y and write down $E(Y)$ and $\mathrm{Var}(Y)$.

(4)

In the case $\lambda = ½$, $n = 50$, use an appropriate approximation to find $P(Y \geq 40)$, and state with a reason whether you would expect your answer to be greater than or less than the true value.

(3)

6. The random variable X is distributed with the geometric probability mass function

$$p(x) = q^{x-1}p \ , \qquad x = 1, 2, 3, \ldots$$

where $0 < p < 1$ and $q = 1 - p$. A random sample $x_1, x_2, \ldots, x_n$ is taken from this distribution.

Write down the likelihood function $L(p)$ based on these data, and show that the maximum likelihood estimate of p is given by

$$\hat{p} = 1/\bar{x}$$

where $\bar{x}$ is the sample mean.

(4)

By using the approximation

$$\mathrm{Var}(\hat{p}) \approx \frac{1}{E\left(-\dfrac{d^2 \ln L}{dp^2}\right)} \ ,$$

or otherwise, show that

$$\mathrm{Var}(\hat{p}) \approx \frac{p^2(1-p)}{n} \ .$$

(4)

[**Note.** You may assume that $E(X) = 1/p$.]

Question 6 continued on next page

A boy counts the number of times that he has to roll a given die in order to obtain a six. His results, summarised, are as follows.

Number of rolls to obtain six x:	1	2	3	4	5	6	7	8	9	10
Frequency f:	7	7	6	4	5	4	2	3	1	2

x	11	13	16	17	20	22	25	33
f	2	2	3	1	2	3	1	1

so that $\sum fx = 448$. Find $\hat{p}$ from these data and use the above result to estimate the variance of $\hat{p}$. Assuming that values of $\hat{p}$ are approximately Normally distributed, calculate an approximate 95% confidence interval for the true probability p of getting a six on a roll of this die.

(6)

Finally, suppose that the die is fair, so that $p = 1/6$. Assuming that values of $\hat{p} = 1/\bar{x}$ are approximately distributed as $N(p, p^2(1 - p)/n)$, where $p = 1/6$, find the approximate probability of the boy obtaining an estimate as low as or lower than that given by the data above.

Comment on your answer.

(6)

7. (a) The continuous random variable X is distributed with probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$. A random sample $X_1, X_2, \ldots, X_n$ is drawn from the distribution. Denote the maximum value of the sample by $X_{\max}$ and the minimum value by $X_{\min}$.

(i) Note that $X_{\max} \le x$ if and only if $X_1 \le x, X_2 \le x, \ldots, X_n \le x$. Hence show that the cdf of $X_{\max}$ is given by

$$F_{X_{\max}}(x) = \left[F(x)\right]^n \quad .$$

(2)

(ii) By noting the condition under which $X_{\min} \ge x$, show that

$$F_{X_{\min}}(x) = 1 - \left[1 - F(x)\right]^n \quad .$$

(2)

(iii) Deduce formulae for the probability density functions of $X_{\max}$ and $X_{\min}$ in terms of $F(x)$ and $f(x)$.

(4)

(b) Suppose that the random variable X has the Pareto density given by

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

Draw a graph of this density and show that the cdf is given by

$$F(x) = 1 - (1+x)^{-\alpha} \quad .$$

Deduce that the median of X is $2^{1/\alpha} - 1$.

(6)

Use result (a)(ii) above to obtain the cdf of the minimum of a random sample of n observations distributed as is X, and verify that this cdf is also of Pareto form but with a different parameter.

(3)

Taking $\alpha = 1$, find the smallest value of n such that the median value of the sample minimum is less than 0.1.

(3)

8. State the model for simple linear regression analysis on one explanatory variable. What are the assumptions usually made regarding the stochastic term?

(3)

Data relating to percentage operating capacity and profits (£) per unit of output are collected for 12 factories producing similar domestic refrigerators in the previous year, as follows.

% operating capacity	50	57	61	68	77	80	82	85	89	91	95	99
profit per unit of output	2.5	4.0	3.1	4.6	7.3	6.2	6.1	11.6	10.0	14.2	16.1	19.5

These data are entered into Minitab, with profit in column 1 (i.e. c1) and % operating capacity in column 2 (i.e. c2), and analysed as shown in the edited output. Use the output to answer the following questions, noting the explanations incorporated in it. c1 is named 'Profits' and c2 is named 'OpCapcty'.

(i) Consider the plot of profits against capacity and the subsequent regression analysis.

(a) Comment on this plot.

(1)

(b) Explain what is meant by the statement R-Sq = 80.3%. How might this have been calculated from the Analysis of Variance immediately following?

(2)

(c) What is the practical meaning of the slope parameter estimate 0.31562 ? Construct a 95% confidence interval (CI) for the slope parameter.

(2)

(d) The subcommand **SUBC> predict …** on page 13 produces the lines starting 'Predicted Values' and ending 'very extreme X values'. Noting that, for example, $-15.799 + (0.31562)(25) = -7.909$ and that the 95% CI is obtained as $-7.909 \pm 2.714\ t_{10,\ 0.025}$, explain the meaning of this section of the output.

(2)

Question 8 continued on next page.
Minitab output follows on pages 13 and 14.

(ii) In the analysis of Model 2, column 3 (i.e. c3) is constructed as $\log_{10}$(profits).

(a) Compare the plot of $\log_{10}$(profits) against % operating capacity with that of profits against % operating capacity.

(2)

(b) Use the regression of $\log_{10}$(profits) on % operating capacity to express profits as a function of % operating capacity.

(2)

(c) Convert the predicted value of $\log_{10}$(profits) for 25% operating capacity, and its CI, to a corresponding prediction and CI for profits. Compare your answers with the corresponding values indicated in the output referred to in part (i)(d) above.

(3)

(iii) Discuss with reasons which of the two regressions you consider provides the better summary of the data. Indicate any limitations which should be borne in mind when using your preferred model.

(3)

Two pages of Minitab output follow

```
MTB > set c1   # profit in £1000s, for 12 factories
DATA> 2.5 4.0 3.1 4.6 7.3 6.2 6.1 11.6 10.0 14.2 16.1 19.5  DATA> end
MTB > set c2   # corresponding % operating capacity for 12 factories as above
DATA> 50 57 61 68 77 80 82 85 89 91 95 99    DATA> end
MTB > name c1 'Profits' c2 'OpCapcty'
```

(i) Analysis of Model 1

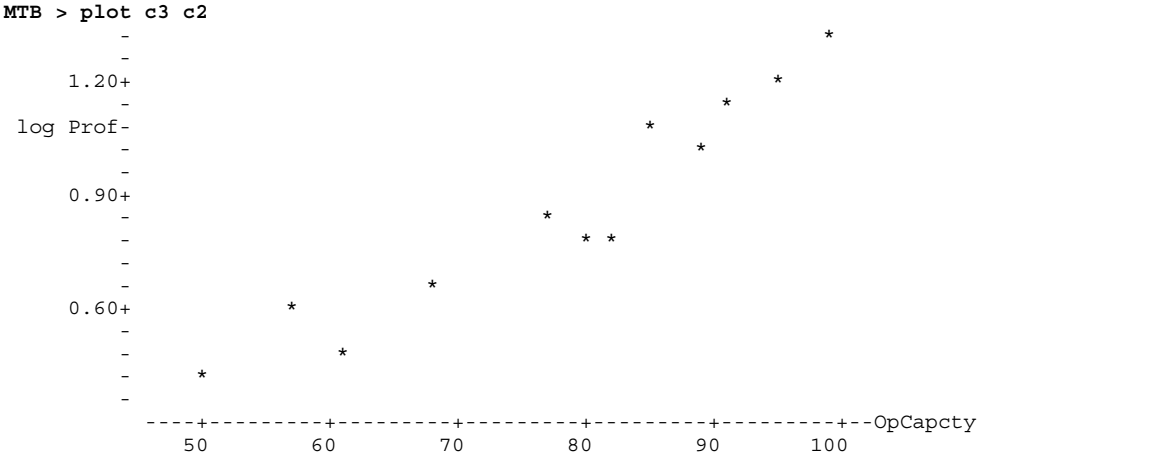
```
MTB > plot c1 c2
          -
          -                                                      *
      18.0+
          -
 Profits  -                                                  *
          -                                              *
          -
      12.0+                                         *
          -
          -                                             *
          -
          -                                   *
       6.0+                                     * *
          -                         *
          -              *   *
          -     *
            ----+---------+---------+---------+---------+---------+--OpCapcty
               50        60        70        80        90       100
```

```
MTB > regress c1 1 c2;
SUBC> residual c3;
SUBC> predict c1 for c2 = 25; SUBC> predict c1 for c2 = 50; SUBC> predict c1 for c2 = 75.

The regression equation is    Profits = - 15.8 + 0.316 OpCapcty

Predictor        Coef       StDev          T        P
Constant      -15.799       3.918      -4.03    0.002
OpCapcty      0.31562     0.04942       6.39    0.000

S = 2.570       R-Sq = 80.3%     R-Sq(adj) = 78.3%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1      269.33      269.33     40.78    0.000
Residual Error    10       66.04        6.60
Total             11      335.37

Predicted Values

     Fit  StDev Fit         95.0% CI
  -7.909      2.714   ( -13.957,  -1.860) XX

     Fit  StDev Fit         95.0% CI
  -0.018      1.563   (  -3.500,   3.464)

     Fit  StDev Fit         95.0% CI
   7.872      0.755   (   6.190,   9.554)
X  denotes use of X values away from the centre
XX denotes use of very extreme X values
```

```
MTB > let c4=c1-c3   # c4 is the fit
MTB > gstd
* NOTE  * Standard Graphics are enabled. Professional Graphics are disabled.
MTB > plot c3 c4          # c3 contains the residuals from regression, c4 is the fit
          -
      c3  -                                                      *
          -
       3.0+
          -   *
          -          *                                           *
          -                                                  *
          -                                             *
       0.0+
          -               *
          -                       *        *
          -
          -                                                  *
      -3.0+                                     *
          -
          -                                       *
            --+---------+---------+---------+---------+---------+----c4
            0.0       3.0       6.0       9.0      12.0      15.0
```

(ii) Analysis of Model 2

```
MTB > let c3=logten(c1)
MTB > name c3 'log Prof'
MTB > plot c3 c2
```




```
MTB > regress c3 1 c2;
SUBC> residual c4;
SUBC> predict c3 for c2 = 25; SUBC> predict c3 for c2 = 50; SUBC> predict c3 for c2 = 75.

The regression equation is    log Prof = - 0.519 + 0.0177 OpCapcty

Predictor        Coef       StDev          T        P
Constant      -0.5185      0.1276      -4.06    0.002
OpCapcty     0.017699    0.001610      10.99    0.000

S = 0.08372     R-Sq = 92.4%     R-Sq(adj) = 91.6%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         1     0.84694     0.84694    120.83    0.000
Residual Error    10     0.07009     0.00701
Total             11     0.91703

Predicted Values

     Fit  StDev Fit         95.0% CI
 -0.0760     0.0884   ( -0.2731,  0.1210) XX

     Fit  StDev Fit         95.0% CI
  0.3664     0.0509   (  0.2530,  0.4799)

     Fit  StDev Fit         95.0% CI
  0.8089     0.0246   (  0.7541,  0.8637)
X  denotes a row with X values away from the centre
XX denotes a row with very extreme X values

MTB > let c5=c3-c4    # c4 contains the residuals from regression, c5 is the fit
MTB > plot c4 c5
```