EXAMINATIONS

April 1999

Subject C — Statistics

Paper One

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Write your surname in full, the initials of your other names and your Candidate's Number on the front of the answer booklet.*
2. *Begin your answers to Parts One, Two and Three on a separate sheet.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 16 questions.*

Graph paper is required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet and this question paper.

In addition to this paper you should have available graph paper, Actuarial Tables and an electronic calculator.

PART ONE

For questions 1–7 indicate in your answer booklet which one of the answers A, B, C or D is correct.

1 A workman charges his customers at the rate of £30 per hour, plus a call-out charge of £40. Over a particular week the mean and standard deviation of the lengths of his jobs are 5hr and 0.5hr respectively.

The mean and standard deviation of the costs (£) incurred by his customers that week are respectively

A 190 and 15
B 190 and 55
C 150 and 7.5
D 150 and 15 [2]

2 In a large collection of life policies, 60% are for male lives, and 15% of the sums assured on these lives exceed £200,000. The percentage of sums assured on female lives which exceed £200,000 is 6%.

A policy is selected from this collection at random. The sum assured on the life concerned is £145,000.

The probability that the selected policy is for a female life is:

A 0.376
B 0.424
C 0.510
D 0.576 [3]

3 For each of a number of independent inquiries at an insurance firm's inquiry desk the probability that it leads to a sale is 0.6. N, the number of inquiries made until the 3^{rd} sale is made, is recorded on a particular day.

The probability that $N = 8$ is

A 0.0464
B 0.1045
C 0.1239
D 0.2787 [3]

4 A simple discrete random variable, X, has probability function given by

$$P(X = 0) = 0.4$$
$$P(X = 1) = 0.6.$$

The coefficient of skewness is:

A -0.048
B -0.098
C -0.20
D -0.41 [3]

5 Suppose that claim sizes of a certain type can be modelled by a normal random variable with mean $\mu = £500$ and standard deviation $\sigma = £100$.

The moment generating function of the difference between two independent claim sizes (£), $M(t)$, is given by:

A $M(t) = \exp(10{,}000t^2)$
B $M(t) = \exp(1000 + 10{,}000t^2)$
C $M(t) = \exp(1000t + 10{,}000t^2)$
D $M(t) = \exp(1000t)$ [3]

6 Let X and Y have joint density function given by

$$f(x,y) = \frac{2}{3}(2x + y) : 0 < x < 1 \, , 0 < y < 1.$$

The conditional density function of X given $Y = y$ is

A $\dfrac{2x + y}{1 + y} : 0 < x < y$

B $\dfrac{2x + y}{1 + y} : 0 < x < 1$

C $\dfrac{2x + y}{2 + y} : 0 < x < y$

D $\dfrac{2x + y}{2 + y} : 0 < x < 1$ [3]

7 It is desired to simulate an observation of the random variable X with probability density function

$$f(x) = \begin{cases} \frac{1}{k}; & 0 \le x \le k \\ 0 & \text{otherwise} \end{cases}$$

A random number r is generated from the uniform distribution over [0, 1]. The following values are then calculated

$$x_1 = rk \quad ; \quad x_2 = k(1-r) \quad ; \quad x_3 = \frac{r}{k}$$

Which of the following statements is correct as regards these values being valid simulated observations of X?

A x_1 and x_2 are, x_3 is not.
B x_1 is. Neither x_2 nor x_3 are.
C x_3 is. Neither x_1 nor x_2 are.
D x_1, x_2 and x_3 all are. [2]

PART TWO

8 In an experiment a coin was tossed six times and the number of heads obtained was recorded. The experiment was repeated 20 times. The following distribution was obtained.

Number of Heads	0	1	2	3	4	5	6
Frequency	2	2	5	5	3	2	1

Calculate the interquartile range of the number of heads obtained per experiment. [3]

9 The number of claims, N, which arise under a group of policies in a year has a Poisson distribution with mean 200. The claim amounts are independent of one another and of N, and have a common distribution which is normal with mean £250 and standard deviation £40.

Let Y be the total claim amount arising from this group of policies in a year.

(i) Show that $E(Y|N) = 250N$ and $V(Y|N) = 1600N$. [2]

(ii) Hence, or otherwise, find the mean and standard deviation of Y. [4]
[Total 6]

10 The probability that an inquiry leads to a sale is 0.7, independently for each inquiry. Over a period of time 200 such inquiries are received. By making a suitable approximation, estimate the probability that N, the number of sales from these inquiries, satisfies

$$124 < N \leq 153$$

[5]

11 A coin with $P(\text{head}) = \theta$ was tossed twice 200 times. The number of heads which appeared each time was noted, with the following results:

Number of heads	0	1	2
Number of occurrences	41	92	67

Find the maximum likelihood estimate of θ, with its approximate standard error.

[5]

12 A random sample of size 5 is taken from an exponential distribution with density

$$f(x;\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}},\ x > 0$$

which has mean θ and variance θ^2.

(i) Show that the sample mean, $\overline{X}$, is unbiased for θ and determine its mean square error. [2]

(ii) The alternative biased estimator $\frac{5}{6}\overline{X}$ is proposed. Determine its mean square error. [3]

(iii) It is suggested that the sample mean, $\overline{X}$, must be the better of the two proposed estimators because it is unbiased. Explain briefly why this is not the case. [2]

[Total 7]

13 A random sample of size 10 is taken from a normal distribution with mean μ and variance σ^2 (both parameters unknown) to test

$$H_0 : \mu = 12 \quad \text{v} \quad H_1 : \mu > 12 .$$

The values of the sample mean and standard deviation are $\bar{x} = 12.940$ and $s = 1.314$ respectively.

Find the approximate probability-value of these sample results. [4]

PART THREE

14 Independent random samples of size n1 and n2 are taken from the normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Let the sample means be $\overline{X}_1$ and $\overline{X}_2$ and the sample variances be S_1^2 and S_2^2.

You may assume that $\overline{X}_i$ and S_i^2 are independent and distributed as follows:

$$\overline{X}_i \sim N\left(\mu_i, \frac{\sigma_i^2}{n_i}\right) \text{ and } \frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1} \;:\; i = 1,2.$$

(i) It is required to construct a confidence interval for $(\mu_1 - \mu_2)$, the difference between the population means.

(a) Suppose that σ_1^2 and σ_2^2 are known. State the distribution of $(\overline{X}_1 - \overline{X}_2)$ and write down a suitable pivotal quantity together with its sampling distribution. Hence write down a 95% confidence interval for $(\mu_1 - \mu_2)$.

(b) Suppose that σ_1^2 and σ_2^2 are unknown but are known to be equal. State the definition of a t_k variable in terms of independent $N(0,1)$ and χ_k^2 variables and use it to develop a suitable pivotal quantity. Hence write down a 95% confidence interval for $(\mu_1 - \mu_2)$. [7]

(ii) It is required to construct a confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$, the ratio of the population variances.

State the definition of an $F_{k,l}$ variable in terms of independent χ_k^2 and χ_l^2 variables and use it to develop a suitable pivotal quantity. Hence obtain a 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$. [5]

(iii) A regional newspaper included a consumer rights article comparing the cost of shopping in "corner shops" and "supermarkets". The researchers investigated the price of a standard "selection" of household goods in a sample of 10 corner shops selected at random from the region, and in a sample of 10 supermarkets selected at random from the region. The data yielded the following values:

	Sample mean	*Sample s.d.*
Corner Shops	£22.55	£1.22
Supermarkets	£19.72	£0.96

(a) Use your result in part (i)(b) to calculate a 95% confidence interval for $(\mu_1 - \mu_2)$, the difference between the population means (1 = corner shops, 2 = supermarkets).

(b) Use your result in part (ii) to calculate a 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$, the ratio of the population variances.

Hence comment briefly on the assumption of equal variances required for the confidence interval in part (iii)(a). [5]

[Total 17]

15 An investment company recently launched a new regular savings scheme under which investors could choose whether to pay £50, £75, £100, or £150 per month to buy units in its new investment. Before the launch a financial analyst was asked for her opinion on the amounts investors would choose. She was of the view that the percentages of investors choosing the various monthly amounts would be as follows:

£50 — 20% £75 — 20% £100 — 50% £150 — 10%

After the launch a random sample of 300 investors in the new savings scheme was examined and the percentage who had in fact chosen the various amounts were as follows:

£50 — 22% £75 — 16% £100 — 56% £150 — 6%

(i) Carry out a test of the goodness-of-fit of the model proposed by the financial analyst, stating clearly the probability–value of the sample results and your conclusion. [5]

(ii) The analyst now suggests that the situation may be more complicated than would be adequately described by the model tested in part (i). In particular she suggests that perhaps a "sex effect" is present, that is to say the pattern of monthly investments differs between men and women.

The 300 investors in the sample are now classified by sex as well as by investment amount and it turns out that the numbers of men who chose to pay £50, £75, £100, or £150 per month are 34, 16, 89, and 4 respectively.

Carry out a contingency table analysis to investigate the presence of a sex effect. [7]

(iii) Discuss briefly (without formally testing) the fit of the analyst's original model to the observed investment pattern for the two sexes separately. [3]

[Total 15]

16 A survey of the durations and fees charged for various continuing education short courses produced the following data.

x duration (hr)	3	5	7	9	11	13	17	19	23	25
y fee (£'00)	2.5	4.0	4.5	4.0	5.0	6.5	6.5	8.0	8.5	10.0

For these data:

$$\sum x = 132 \ , \ \sum y = 59.5 \ , \ \sum xy = 942.5 \ , \ \sum x^2 = 2258 \ , \ \sum y^2 = 404.25$$

It is thought that a suitable model to represent the data is

$$Y = \alpha + \beta x + \varepsilon$$

where Y is the fee charged, x is the duration and e is a $N\left(0, \sigma^2\right)$ error term.

(i) Plot the data and comment on the suitability of the above model. [2]

(ii) Derive formulae for $\hat{\alpha}$ and $\hat{\beta}$, the least squares estimators of α and β respectively. [4]

(iii) Give an economic interpretation of the parameters α and β in the light of the given data. [2]

(iv) Calculate the values of $\hat{\alpha}$ and $\hat{\beta}$ and hence derive a 95% confidence interval for

(a) β , and

(b) a , which is the average fee for courses of zero hours duration. [9]

(v) Briefly explain why it is reasonable to estimate the fee for a course of about 13 hours duration, but it would be inappropriate to repeat the exercise for one of about 30 hours.

[2]

[Total 19]