Full marks can be obtained for correct answers to
**five questions**

Data provided: New Cambridge Elementary Statistical Tables by D.V. Lindley and W.F. Scott.

**Some Useful Formulae**

1) For any two events A and B

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$,

$P(A \cap B) = P(A/B)\,P(B)$,

$P(A \cap \overline{B}\,) = P(A) - P(A \cap B)$.

2) If X has a Binomial distribution with parameters n and p

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (x = 0, 1, ..., n),$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

and for each integer x ≥ 1, x! = x(x-1)(x-2) ... 1, 0! = 1.

Also, E(X) = np, V(X) = np(1-p).

3) If X has a Poisson distribution with mean λ

$$P(X = x) = \frac{\lambda^x}{x!} \exp(-\lambda) \qquad (x = 0, 1, ...),$$

and

$$E(X) = \lambda,\ V(X) = \lambda.$$

Moreover, under suitable conditions,

$$P(a \le X \le b) = \Phi(\beta) - \Phi(\alpha),$$

where $\beta = \frac{(b + 0.5 - \lambda)}{\{\lambda\}^{1/2}}$, $\alpha = \frac{a - 0.5 - \lambda}{\{\lambda\}^{1/2}}$

and $\Phi(z)$ denotes the area to the left of z for a standard Normal distribution.

1. Haemoglobin levels (in grams per millilitre) of lung cancer patients in a military hospital were recorded as follows to the nearest one decimal place:

| | | | | | | |
|---|---|---|---|---|---|---|
| 13.5 | 15.6 | 16.3 | 12.3 | 13.1 | 14.2 | 12.4 |
| 11.3 | 14.0 | 14.6 | 13.6 | 14.8 | 12.7 | 10.9 |
| 11.0 | 11.4 | 15.0 | 10.1 | 15.4 | 11.3 | 10.7 |
| 14.6 | 13.5 | 15.0 | 12.1 | 12.0 | 14.2 | 11.4 |
| 15.0 | 13.3 | 13.2 | 9.1 | 16.9 | 14.2 | 15.0 |
| 13.6 | 14.8 | 11.4 | 14.8 | 15.7 | 13.5 | 13.5 |

a) Group the data into a frequency distribution with class intervals defined as 8.95 - 9.95, 9.95-10.95, 10.95-11.95, 11.95-12.95, 12.95-13.95, 13.95-14.95, 14.95-15.95, 15.95-16.95.

[7 marks]

b) Calculate the mean and standard deviation of the data *from your frequency distribution*.

[8 marks]

c) Without carrying out a formal hypothesis test, compare visually the observed proportion of patients with haemoglobin levels less than 10.95 grams/millilitre with that which would be implied by a Normal distribution with the mean and standard deviation calculated as in b) above. [5 marks]

2. The data presented below are two basic stem and leaf plots of serum cholesterol levels (in mg/litre) for men from two socio-economic groups in Guatemala: 49 low income rural men and 45 high income urban men. The leaf unit is 1 mg/litre and 9|5 for example represents 95 mg/litre.

**Table: Basic stem and leaf plots for serum cholesterol levels (in mg/litre) for men from two different socio-economic groups in Guatemala**

Example 9|5 represents 95 mg/l

| **Rural** | |
|---|---|
| 9 | 5 |
| 10 | 8, 8 |
| 11 | 4, 5 |
| 12 | 4, 9, 9 |
| 13 | 1, 1, 5, 6, 6, 9 |
| 14 | 0, 2, 2, 3, 3, 4, 4, 5, 6, 8 |
| 15 | 2, 2, 5, 7, 8, 8 |
| 16 | 2, 5, 6 |
| 17 | 1, 2, 3, 4, 5 |
| 18 | 0, 1, 9 |
| 19 | 2, 4, 7 |
| 20 | 4 |
| 21 | |
| 22 | 0, 3, 6 |
| 23 | 1 |

| **Urban** | |
|---|---|
| 13 | 3, 4 |
| 14 | |
| 15 | 5 |
| 16 | |
| 17 | 0, 5, 9 |
| 18 | 1, 4, 8, 9 |
| 19 | 0, 6, 7, 9 |
| 20 | 0, 0, 1, 1, 4, 5, 5, 5, 6 |
| 21 | 4, 7 |
| 22 | 2, 2, 7, 7, 8 |
| 23 | 4, 4, 6, 9 |
| 24 | 1, 2, 4, 9 |
| 25 | 2 |
| 26 | |
| 27 | 3, 9 |
| 28 | 4, 4, 4 |
| | |
| HI | 330 |

Obtain a comparative Box plot of the data. [15 marks]

Comment on your results. [5 marks]

3. Explain what is meant by

i) a random experiment [2 marks]

ii) elementary outcomes of a random experiment [1 mark]

Two standard fair dice, one coloured blue, the other red, and each with faces numbered 1, 2, ..., 6 are thrown. If a 'double' is obtained (that is, the numbers shown on the uppermost faces of the two dice are both the same) then a third standard fair dice, coloured green, is thrown.

Let (i, j) denote the outcome that the number on the uppermost face of the blue die is i and that on the uppermost face of the red die is j, $i \neq j$, and let (i, i, k) denote the outcome that the numbers on the uppermost faces of the blue and red dice are both i and that on the green die is k (i, j, k = 1, ..., 6).

List the elementary outcomes of the experiment. [3 marks]

Explain why

a) for i, j = 1, ..., 6, with $i \neq j$,

P(outcome (i, j) is observed) = $\frac{1}{36}$; [1 mark]

b) P(Green die is thrown) = $\frac{1}{6}$; [2 marks]

c) for i, k = 1, ..., 6,

P(outcome (i, i, k) is observed) = $\frac{1}{216}$. [3 marks]

Deduce that not all elementary outcomes of the experiment are 'equally likely'.

[1 mark]

***Question 3 continued overleaf***

***Q3 contd***

The total score obtained at the end of the experiment was determined, where the total score is defined as the sum of the numbers occurring on the uppermost faces of the blue and red dice, if a double was not obtained, and of those occurring on the faces of the blue, red and green dice, if a double was obtained.

Find the probability that

d) score of 18 was obtained; [1 mark]

e) score was 4 given that a double was not obtained; [2 marks]

f) score was 4 given that a double was obtained; [2 marks]

g) score was 4. [2 marks]

4. In a certain area, two traffic wardens, A and B, say, issue tickets for parking violations. The number of tickets issued by A in an hour may be modelled by a Poisson distribution with a mean of 5 and that by B may also be modelled by a Poisson distribution, but with a mean of 8 tickets per hour.

Find the probability that

a) A issues at least one parking ticket in an hour; [2 marks]

b) the combined total number of parking tickets issued by A and B in one hour is at least 5. [5 marks]

[**N.B.** You may assume without proof the standard result that if the random variables $X$ and $Y$ are independent, $X$ is Poisson with mean $\lambda$, $Y$ is Poisson with mean $\mu$, then $X+Y$ is Poisson with mean $\lambda+\mu$.]

Suppose that A and B each issue tickets for six hours per day but the tickets are in fact issued for a total of 10 hours per day, and two part-time traffic wardens are employed for the remaining 4 hours. The numbers of tickets issued by each of these two part-time wardens also follow independent Poisson distributions, but each with a mean of 3 tickets per hour, and independent of the numbers of tickets issued by A and B. On using an appropriate Normal approximation to the Poisson distribution, find

c) the probability that more than 125 tickets are issued per day; [6 marks]

d) an integer, $k$, such that the probability of the number of tickets issued per day being less than or equal to $k$ is no more than 0.01. [7 marks]

[**N.B**. If Z is Normally distributed with mean 0 and variance 1

$$P(Z < -2.3263) = 0.01]$$

5. The proportion, *X*, of alcohol in a certain compound may be considered as a random variable with the probability density function

$$f(x) = 20x^3 (1 - x), \quad 0 < x < 1.$$

a) Evaluate $P\left(X \leq \frac{2}{3}\right)$, correct to 5 decimal places. [6 marks]

b) Find the expected value of *X*. [3 marks]

c) Suppose that the selling price of the above compound depends on the alcohol content. Specifically, if $\frac{1}{3} < X < \frac{2}{3}$, the compound sells for £20.00/litre; otherwise it sells for £10.00/litre. The cost of producing the compound is £12.00/litre.

i) Find the probability function of the net profit per litre. [9 marks]

ii) Evaluate the expected net profit per litre. [2 marks]

6. In a study of the gender distribution of children in families with 3 children, the numbers of families with *k* male children were recorded for *k* = 0, 1, 2, 3.

For a randomly selected family with 3 children, let *X* denote the random variable that *k* children are male. Explain the conditions under which *X* follows a Binomial distribution with parameters 3 and *p*, i.e.

$$P(X = k) \;=\; \binom{3}{k} \quad p^k (1-p)^{3-k} \qquad (k = 0, 1, 2, 3). \qquad (1)$$

[4 marks]

The following table was constructed from the data collected by K. Pakrasi and A. Halder (1971) in a study of the gender distribution of children in India and it gives the numbers of families with k male children, k = 0, 1, 2, 3, in a sample of 19,788 families with 3 children in urban areas of India.

**Table**

**Number of male children in families with 3 children living in urban areas of India**

| **No of male children (*k*)** | **No of families ($0_k$)** |
|---|---|
| 0 | 1990 |
| 1 | 7134 |
| 2 | 8034 |
| 3 | 2630 |
| **Total** | **19788** |

Find an estimate, $\hat{p}$, say, of the probability that a child at birth is male, correct to 5 decimal places. [4 marks]

Fit the binomial distribution given in (1) above to this data and test the goodness of fit.

[9 marks]

***Question 6 is continued overleaf***

***Q6 contd***

Explain giving reasons why a Binomial distribution may not be an appropriate probability model for the number of male children in a family whether or not living in urban areas of India, and suggest any special factors that may make the binomial model particularly inappropriate for families living in the urban areas of India. [3 marks]

7. The data given below were collected in the US in 1989, 1990 and 1991 and are based on responses given by a simple random sample of men and women from the US population 18 years of age or older to a self administered questionnaire designed to obtain information on current patterns of adult sexual behaviour in the general population. Respondents were asked to report the number of male and female sexual partners they had had since they were 18 years of age (referred to here as 'lifetime partners')

A) In the Table given below, the data for 1989, 1990 and 1991 are aggregated and the respondents are classified according to gender and according to whether they reported less than 10 opposite-sex lifetime partners or 10 or more such partners.

**Table**

**Number of opposite-sex lifetime sexual partners classified according to gender and sexual activity.**

| | **Sexual activity** | | |
|---|---|---|---|
| **Gender** | **Fewer than 10 partners** | **Ten or more partners** | **Total** |
| Male | 3240 | 16471 | 19711 |
| Female | 3429 | 2717 | 6146 |
| **Total partners** | **6669** | **19188** | **25857** |

A puzzling anomaly is that whereas in the entire population the number of lifetime female sexual partners reported by men may be expected to equal the number of lifetime male partners reported by females, in the sample the former far exceeds the latter. To investigate this anomaly further, test the hypothesis that the variables 'Gender' and 'Sexual Activity' are independent. [8 marks]

***Question 7 is continued overleaf***

***Q7 contd***

B) It may be gleaned from the Table above that much of the gender discrepancy in the reported sexual activity may originate in the highly active group with 10 or more lifetime partners. To investigate this observation further, consider only the data for the sexually less active group with fewer than 10 lifetime partners. For this group only, test the hypothesis that there are no gender differences in the reported sexual activity, that is, in the entire population the number of female lifetime sexual partners reported by sexually less active males and the number of male lifetime partners reported by sexually less active females are equal. [8 marks]

C) Relate the result obtained in B) above to that in A) and comment on what additional information may be collected to resolve the anomaly described in A) above. [4 marks]