Instructions to Candidates:

1. Answer any 4 questions.
2. Questions may be answered in any order but your answers must show the question number clearly.
3. Always start a new question on a fresh page.
4. All questions carry equal marks

This question paper contains 5 questions and 8 pages.
QUESTION 1: (25 MARKS)

(a) Distinguish carefully between stratified sampling, cluster sampling and quota sampling, stating the benefits and drawbacks of each method. \( (12 \text{ marks}) \)

(b) A country is divided into regions and in each region there are both urban and rural areas. A survey is to be undertaken in which adults are to be interviewed. Devise a sampling scheme which is a combination of stratified, cluster and quota sampling. \( (4 \text{ marks}) \)

(c) A village has 506 inhabitants, listed by name and address in a register. Explain in detail how to take a systematic sample of about 50 inhabitants from this list. State the benefits and drawbacks of this method. \( (5 \text{ marks}) \)

(d) One thousand people were selected to participate in a face-to-face survey. Of the thousand, 725 were at home at the first attempt to contact them but 52 of these refused to take part. One call-back was made to each person missed at the first attempt, and 30 of these people were not at home at the time of call. Of those that were at home, 49 refused to take part in the survey. Calculate the response rate at the first contact, the response rate at the second contact and the overall response rate. \( (4 \text{ marks}) \)
QUESTION 2: (25 MARKS)

(a) A blended wine is intended to comprise two parts of Sauvignon to one part of Merlot. The amounts dispensed to make up a nominal 75 cl bottle of this wine are \( X \) cl of Sauvignon and \( Y \) cl of Merlot, where \( X \) and \( Y \) are assumed to be independent Normally distributed random variables with respective means 52 and 26 cl and respective variances 1 and 0.5625. Find the probability that the actual volume of wine dispensed into a bottle is less than the nominal volume.

(4 marks)

(b) The continuous random variable \( X \) has probability density function (pdf)

\[
f_X(x) = \frac{1}{2\theta}, \quad -\theta \leq x \leq \theta.
\]

Show that the cumulative distribution function of \( X \) is

\[
F_X(x) = \frac{\theta + x}{2\theta}, \quad -\theta \leq x \leq \theta.
\]

Deduce an expression for \( P(X > x) \).

(5 marks)

(c) The Poisson random variable \( X \) with parameter \( \lambda > 0 \) has probability mass function

\[
p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots.
\]

i. Show that, for integer \( x \geq 0 \),

\[
p_X(x + 1) = \frac{\lambda}{x + 1} p_X(x),
\]

(3 marks)

ii. Show that \( E(X) = \text{Var}(X) = \lambda \).

(6 marks)
(d) An office has two computer systems, one of Type A and one of Type B. The numbers of breakdowns per day on these systems $X$ and $Y$ say, have independent Poisson distributions with respective means 2 and 0.5.

i. Find the conditional probability that if there is exactly one breakdown on a given day, then it is the Type A system that fails. \hspace{1cm} (2 marks)

ii. Find the probability that on a given day there are more than 2 breakdowns. \hspace{1cm} (2 marks)

iii. The office is one of 50 run by a large company. The offices are each equipped with one Type A system and one Type B system, which function independently in the way described above. Write down the distribution of $T$, the total number of breakdowns occurring in the 50 offices on any given day. Use a suitable approximation to estimate $T_{0.95}$, the number of breakdowns per day which will be exceeded on at most 5% of days. \hspace{1cm} (3 marks)
QUESTION 3: (25 MARKS)

(a) The number of eggs laid by a breeding female of a certain species of sea bird has a Poisson distribution with mean $\lambda$ (> 0) and is independent for different birds. An ornithologist wants to estimate $\lambda$ and examines a number of possible nests. She can only be sure that the nest belongs to the correct species if there is at least one egg in it. The number of eggs in the $i^{th}$ nest identified as belonging to this species is denoted by $X_i$; by the end of the day, she has identified $n$ (> 0) such nests. (Note that $X_i > 0$ for $i = 1, 2, \ldots, n$.)

i. Obtain an expression of $P(X_i = k)$ for $k = 1, 2, 3, \ldots$. (3 marks)

ii. Find an equation for determining $\hat{\lambda}$, the maximum likelihood estimator of $\lambda$ based on $X_1, X_2, \ldots, X_n$. (6 marks)

(b) 12 electric components of Type A and 10 of Type B were tested for reliability. The lifetimes (in hours) of these components were recorded as follows:

<table>
<thead>
<tr>
<th>Type A</th>
<th>162</th>
<th>153</th>
<th>144</th>
<th>156</th>
<th>165</th>
<th>155</th>
<th>172</th>
<th>141</th>
<th>158</th>
<th>156</th>
<th>166</th>
<th>147</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type B</td>
<td>163</td>
<td>158</td>
<td>155</td>
<td>164</td>
<td>153</td>
<td>173</td>
<td>149</td>
<td>171</td>
<td>157</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test whether there is a significant difference in reliability between the two types (at 10% level of significance). (8 marks)

(c) For 1000 drivers sampled, the number of accidents of each driver in the past three years was recorded below:

<table>
<thead>
<tr>
<th>No. of Accidents</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of drivers</td>
<td>100</td>
<td>267</td>
<td>311</td>
<td>208</td>
<td>87</td>
<td>23</td>
<td>4</td>
</tr>
</tbody>
</table>

Test whether the Poisson distribution with an unknown mean $\lambda$ is suitable to model the data, using the chi-square test at 5% level of significance. (8 marks)
QUESTION 4: (25 MARKS)

(a)

i. Briefly explain the principles of randomisation and replication, in the context of a completely randomised experimental design. Your answer should make reference to an example. (6 marks)

ii. Write down the model equation for a completely randomised design having equal numbers of replicates in all treatment groups, defining all the symbols that you use. (5 marks)

(b) The table below shows breaking strains, y kg, of 5 samples of steel wire from each of 4 suppliers, A, B, C and D.

<table>
<thead>
<tr>
<th>Supplier</th>
<th>Breaking strains</th>
<th>$\sum y$</th>
<th>$\sum y^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>137, 133, 136, 134, 130</td>
<td>670</td>
<td>89810</td>
</tr>
<tr>
<td>B</td>
<td>136, 134, 140, 133, 137</td>
<td>680</td>
<td>92510</td>
</tr>
<tr>
<td>C</td>
<td>134, 142, 137, 143, 139</td>
<td>695</td>
<td>96659</td>
</tr>
<tr>
<td>D</td>
<td>144, 140, 143, 147, 141</td>
<td>715</td>
<td>102275</td>
</tr>
</tbody>
</table>

i. Carry out an analysis of variance to test for differences between the mean breaking strains of the steel wire manufactured by these four suppliers. State your conclusions clearly. (12 marks)

ii. State the assumptions made in your analysis. (2 marks)
QUESTION 5: (25 MARKS)

(a) The Devon Motor Racing Grand Prix takes place every five years. Winning average lap speeds (in miles per hour) in the last nine events are shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed y</td>
<td>109</td>
<td>114</td>
<td>116</td>
<td>117</td>
<td>114</td>
<td>127</td>
<td>131</td>
<td>138</td>
<td>141</td>
</tr>
</tbody>
</table>

You are given that

\[ \bar{x} = 1985, \quad \sum (x - \bar{x})^2 = 1500, \quad \sum y = 1107, \quad \sum y^2 = 137233, \quad \sum (x - \bar{x})y = 1200. \]

i. Plot these data and comment on their suitability for simple linear regression analysis. \( (4 \text{ marks}) \)

ii. Fit a simple linear regression model and state its equation. Also compute the total sum of squares and regression sum of squares for this regression, and deduce the error mean square. \( (7 \text{ marks}) \)

iii. It is later noted that driving conditions in 1985 were affected by a freak thunderstorm which caused partial flooding of the track. The 1985 values were therefore omitted and the regression was recalculated. Results are shown in the computer output below. Compare this analysis with your own results and say with reasons which you regard as the more satisfactory. \( (4 \text{ marks}) \)

The regression equation is \( y = -1464 + 0.800x \)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1463.87</td>
<td>95.60</td>
<td>-15.31</td>
<td>0.000</td>
</tr>
<tr>
<td>x</td>
<td>0.80000</td>
<td>0.04816</td>
<td>16.61</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\( S = 1.86525, \quad R-Sq = 97.9\%, \quad R-Sq(adj) = 97.5\% \)

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>960.00</td>
<td>960.00</td>
<td>275.93</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>6</td>
<td>20.87</td>
<td>3.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>980.87</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
iv. Use the analysis of part (iii) to obtain point estimates of

A. the expected winning speed in 1985, (2 marks)
B. the expected winning speed in 2010, (1 marks)
C. the time by which a winning speed of 160 mph might be expected. (2 marks)
D. Mention any reservations you might have about your answers. (2 marks)

(b) In the multiple linear regression model

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad i = 1, \ldots, n, \]

what assumptions are usually made about the residual (error) terms \( e_1, \ldots, e_n \)? (3 marks)

***END OF QUESTION PAPER***