1. (a) A town is protected from flood by a reservoir dam designed to withstand a 100 year flood i.e. the probability in a given year that the dam overflows is 0.01. The reservoir is also located in an active seismic region. During an earthquake, there is a 30% chance that the dam will be damaged and the town flooded. The occurrence of floods due to an earthquake or dam overflow are assumed to be independent. Also, the occurrence of floods due to different earthquakes are assumed to be independent.

   (i) If one earthquake occurs in a year, what is the probability that the town will be flooded that year?
   (ii) If there are two earthquakes in a year, what is the probability that neither causes a flood?
   (iii) What is the probability that the town is flooded if, from any cause, two earthquakes occur in a year?

   (b) In Bayestown, 3% of the population are guilty of a crime. The police in Bayestown use a truth serum to decide whether someone is guilty of a crime.
However, the serum is not 100% reliable. When given the serum, a guilty person confesses to a crime 95% of the time and an innocent person maintains his innocence 99% of the time; i.e. 5% of guilty people are found innocent and 1% of innocent people are found guilty.

(i) What proportion of the population in Bayestown will admit to a crime if given the serum?
(ii) A crime is committed and a suspect is arrested. He is given the serum and admits to the crime. What is the probability that he is actually guilty?

(c) A central fire alarm system detects fires in an office building from a large number of smoke alarms. Each year, the probability of a particular smoke alarm going off is very small. The total number of alarms raised in the building each year is Poisson distributed with a mean of 4.

(i) Write down the probability distribution for the total number of alarms raised.
(ii) Give a reason why the Poisson distribution might be a reasonable model for the number of alarms raised.
(iii) What is the probability that no alarms will be raised in a year?
(iv) What is the probability that three or more alarms will be raised in a year?

2. An internet provider company is analysing the response time from its server. At 20 different times during the day, it measured the response times of its server.

(a) These data were analysed by Data Desk, which produced the following output:

<table>
<thead>
<tr>
<th>Summary of Response times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td>IntQRange</td>
</tr>
<tr>
<td>Std. Error</td>
</tr>
</tbody>
</table>

(i) Explain each of the terms: Mean, Median, Variance, Range, IntQRange and Std. Error.
(ii) What is the coefficient of variation of these data?
(b) Form a 95% confidence interval for the variance in response times.

(c) The company’s main competitor claims a mean response time of 5 seconds. The company wants to be able to claim a shorter mean response time, and does a statistical test on Data Desk, which produces the following output:

<table>
<thead>
<tr>
<th>t-Test of Individual μ’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Alpha Level 0.05</td>
</tr>
<tr>
<td>Ho: μ = 5 Ha: μ &lt; 5</td>
</tr>
<tr>
<td>Test Ho: μ(response times) = 5 vs Ha: μ(response times) &lt; 5</td>
</tr>
<tr>
<td>Sample Mean = 3.575, t-Statistic = -2.771 w/19 df</td>
</tr>
<tr>
<td>Reject Ho at Alpha = 0.05</td>
</tr>
<tr>
<td>p = 0.0061</td>
</tr>
</tbody>
</table>

(i) Say how the test statistic value of −2.771 was calculated (you do not have to carry out the actual calculations).
(ii) Explain the meaning of the value “p=0.0061” that appears at the end of the Data Desk output.
(iii) Can the company claim to have a shorter mean response time than its competitor? Explain your reasoning.

3. (a) A PC manufacturer is trying to choose between zip drives from two different suppliers. In an effort to choose between them, it conducts a series of tests. Twelve of the zip drives from the first supplier and ten from the second are taken and the same information copied onto each drive. The number of write errors for each drive is then measured. The twelve drives from the first supplier had a mean of 4.5 write errors, with a standard deviation of 1.3. The ten drives from the second supplier had a mean of 3.9 write errors and a standard deviation of 1.5.

(i) The manufacturer wants to know if there is any significant difference between the mean number of write errors that each drive makes. Write down the null and alternative hypotheses of an appropriate statistical test.
(ii) The reviewer calculates a t-test statistic value of 1.01. Say how this value was calculated (you do not have to do the actual calculations).
(iii) What do you conclude from the test? Explain your reasoning.
(iv) What assumption is being made about the distribution of the sample means of the number of write errors from each supplier? Explain why this assumption can be made.
(b) An alternative test for the two makes of zip drive is as follows. One drive from each supplier is taken and the same set of data is loaded into them under 8 different operating conditions (such as temperature, humidity and CPU load). The number of write errors that occurred under each set of operating conditions is recorded as follows:

<table>
<thead>
<tr>
<th>Drive 1</th>
<th>3</th>
<th>4</th>
<th>0</th>
<th>3</th>
<th>5</th>
<th>4</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive 2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

(i) The manufacturer still wants to know if there is a difference in the number of errors made by each zip drive. In this case, it chooses to do a paired t-test on the above data. Explain why a paired t-test is appropriate in this case.

(ii) The test statistic value for the paired t-test is 2.39. Say how this value has been calculated (you do not have to do the actual calculations).

(iii) What do you conclude from this test? Explain your reasoning.

4. (a) A study of the occurrence of faults caused by a company’s payroll software has collected data on the number of faults each month. Over 38 months, the number of faults each month were recorded and occurred with the following frequencies:

<table>
<thead>
<tr>
<th>No. of faults</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

The mean number of faults over the 38 months was 1.275.

A theory as to how errors in complex software systems arise predicts that the number of errors per month should be Poisson distributed.

(i) If the number of faults is Poisson distributed, in how many of the 38 months would you expect to see zero faults?

(ii) A manager at the company wants to test if the above data are consistent with a Poisson distribution. She conducts a $\chi^2$ test and
arrives at a test statistic value of 0.828. Explain how this value was calculated (you do not have to do the actual calculations).

(iii) What do you conclude from the test statistic value? Explain your reasoning.

(b) In a genetic study, 30 individuals are classified according to whether they possess a certain genetic aberration and whether one, two or neither of their parents was a carrier. The following contingency table was constructed from the classification:

<table>
<thead>
<tr>
<th></th>
<th>Both parents</th>
<th>One parent</th>
<th>Neither parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aberration</td>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>No aberration</td>
<td>0</td>
<td>5</td>
<td>14</td>
</tr>
</tbody>
</table>

(i) The study wants to test the hypothesis that the aberration is associated with the number of parental carriers. A $\chi^2$ test is to be used. Write down the null hypothesis for this test.

(ii) If the null hypothesis is true, how many of the 30 individuals would you expect to see with the aberration and with both parents as carriers?

(iii) The test statistic value is found to be 9.55. Explain how this value was calculated (you do not have to do the actual calculations).

(iv) What do you conclude from the value of the test statistic? Explain your reasoning.

5. (a) An engineer is trying to measure the mean time to failure of a component. He tests four of the components and they fail after 5, 4.3, 6.2 and 5.1 days.

(i) Calculate the mean and standard deviation of the four failure times.

(ii) The company wants to be 95% confident that the mean time to failure of the component is known to within 0.2 days. How many more components must be tested to achieve this goal?

(b) A supercomputer development company builds parallel computers, consisting of several ‘nodes’, each node consisting of a CPU, memory and links to other nodes. The company is investigating the effect of the number of nodes on the speed at which a certain numerical calculation can be completed.
It builds computers consisting of 2, 4, 8, 10 and 15 nodes, and measures the time to complete the calculation as follows:

<table>
<thead>
<tr>
<th>Nodes</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (secs)</td>
<td>112</td>
<td>80</td>
<td>40</td>
<td>30</td>
<td>18</td>
</tr>
</tbody>
</table>

Using Data Desk, the company statistician performs a linear regression, with the following output:

Dependent variable is: Time

R squared = 88.2%  R squared (adjusted) = 84.2%  s = 15.5  with 5 - 2 = 3 degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>5367.37</td>
<td>1</td>
<td>5367.37</td>
<td>22.3</td>
</tr>
<tr>
<td>Residual</td>
<td>720.634</td>
<td>3</td>
<td>240.211</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>111.821</td>
<td>13.69</td>
<td>8.17</td>
<td>0.0038</td>
</tr>
<tr>
<td>Nodes</td>
<td>-7.15649</td>
<td>1.514</td>
<td>-4.73</td>
<td>0.0179</td>
</tr>
</tbody>
</table>

(i) Why is “time” the dependent variable for this regression?
(ii) What is the least squares regression line for time on number of nodes? Explain in what sense this is the best straight line that can be fitted to the data.
(iii) The company wants to know if the number of nodes has a significant effect on the computation time. It conducts a hypothesis test to see if the slope of the regression line is different from 0. What is the result of that test?
(iv) What is the prediction for the time to complete the calculation for a 50 node computer? Give reasons why this prediction should be treated with caution.

6. (a) Define a ‘pseudo-random’ number. Describe one method by which a computer generates pseudo-random numbers and discuss how these numbers differ from truly random numbers.

(b) A computer generates the following sequence of four pseudo-random numbers: 0.11, 0.65, 0.98, 0.54. Use these numbers to generate, using the inverse distribution method, four numbers from the binomial distribution with n=3 and p=0.6. The probability distribution for this binomial is given in the table overleaf.
\[
\begin{array}{|c|c|c|c|c|}
\hline
x & 0 & 1 & 2 & 3 \\
\hline
P(x) & 0.064 & 0.288 & 0.432 & 0.216 \\
\hline
\end{array}
\]

(c) Develop a rejection algorithm to generate from the probability distribution with density function:

\[
f(x) = \begin{cases} 
0.5 \sin(x), & \text{if } 0 \leq x \leq \pi \\
0, & \text{otherwise}
\end{cases}
\]

Note that the maximum value of \(f(x)\) in the range \(0 \leq x \leq \pi\) is 0.5.

SECTION B

7. Define the finite precision floating point system and describe six features which arise from using this system.

8. Describe fully the Romberg Algorithm for numerical integration. It is not necessary to derive any formula used.