COMP116 – Work Sheet Six

Associated Module Learning Outcomes

1. Basic understanding of the range of techniques used to analyse and reason about computational settings.

Statistics & Data Analysis

Q1: Basic Statistical Measures

The first question concerns the **three** samples given below:

$$\begin{array}{lcl} A & = & \{10,\ 10,\ 25,\ 80,\ 2,\ 32,\ 100\} \\ B & = & \{5,\ 90,\ 15,\ 40,\ 35\} \\ C & = & \{1,\ 2,\ 95,\ 90,\ 5,\ 85,\ 8,\ 10,\ 37\} \end{array}$$

a. What are the **average** value(s) for each case (i.e. $E[A]$, $E[B]$ and $E[C]$).

b. Similarly what are the **median** values for A, B and C.

c. Suppose these three samples arose as the outcome of collections of 7 (sample A), 5 (sample B) and 9 (sample C) student marks represented by different groups taking different class tests.

   1. Does anything seem "unreasonable" in the respective performances of the three groups with respect to the different test papers used?
   2. Of the three different papers which, in your view, seemed to be the "fairest" assessment?
   3. Give a more formal justification of your answer to (2) by computing the **variance** within each of the three samples.

d. Suppose, instead of analyzing the three data sets separately these are collected together into a single population of 21 members. What would be the average, median and variance for the resulting collection?

e. The three separate tests that had led to the data in A, B, C are replaced using a single test of all 21 candidates. This results in the outcome shown in Table 1.

Table 1: Test outcomes

Score	# Candidates Achieving	Score	# Candidates Achieving
10	1	55	2
15	2	64	2
20	2	70	3
35	2	85	3
45	2	90	2

1. What are now the median, average, and variance of the outcome?
2. Do you consider (on the basis of your answer to (1) and the overall performance) that the new test is "fairer", "worse" or "just as bad/good" as the three separate tests it replaced?

Q2: Data Analysis

This question concerns applying **Linear regression** methods (discussed in Lectures and discussed in Section 6.12, pages 331–349 of the course textbook).

An experimental study reports a set of **ten** observations presented in Table 2.

Table 2: Experiment observation

Value for x	Outcome y
1	0.19
2	0.483
3	0.64
4	0.76
5	0.95
6	1.371
7	1.33
8	1.52
9	2.016
10	1.9

a. Using **Linear Least Squares** (course textbook page 341) find the best fit **line** for these data.

b. Similarly (as described on pages 344–5) find the best fit function of the form $f(x) = \alpha x^{\beta}$ for these data.

c. Comparing the two functions found against the actual data which of the two looks like a "better match"?

d. How might it be possible to justify your answer to (c) using **only** the experimental data and the two functions discovered? (that is without relying on subjective opinions).