COMP116 – Work Sheet Six – Solutions

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?

Solutions

a.

$$\begin{aligned} E[A] = \frac{10+10+25+80+2+32+100}{7} &= 37 \\ E[B] = \frac{5+90+15+40+35}{5} &= 37 \\ E[C] = \frac{1+2+95+90+5+85+8+10+37}{9} &= 37 \end{aligned}$$

b. The median values are those in the middle of sorted (ascending or descending order doesn't matter since samples have odd numbers of elements). For A this is the middle value from $< 2, 10, 10, 25, 32, 80, 100 >$, i.e. 25. For B the middle value in $< 5, 15, 35, 40, 90 >$ i.e. 35. Finally with C the middle element of $< 1, 2, 5, 8, 10, 37, 85, 90, 95 >$, i.e 10.

c. 1. All 3 papers report the same average performance with A and B having "similar" although differing medians. There may be a case for seeing (C) as unusual given the lower median but this is unclear.

2. Superficially (B) appears to have been the fairest: all but one mark (90) are in a similar performance range ($[5, 40]$). Case (A) has extremes of very low ($\{2, 10\}$) to extremely high ($\{80, 100\}$). Similarly (C), in addition to the skewed median has extremes at both ends: $\{1, 2, 5, 8\}$ against $\{85, 90, 95\}$.
3. For (A) the variance is found to be ~ 1238.57; for (B) it is 866. Finally (C) exhibits a variance of ~ 1510.22. Overall, (B) has same average as the other two samples, median similar to (A) and the marks are least spread out (i.e. it has the lowest variance of the three samples). While this provides a little support for regarding (B) as the "fairest" assessment, it is, however, based on the smallest sample size.

d. The average will be unchanged. All three samples had the same average so the result of aggregating these will just be

$$\frac{37|A| + 37|B| + 37|C|}{|A| + |B| + |C|} \quad = \quad 37$$

The median value over the entire sample of 21 is 25, the 11th lowest score (10 lower are $< 1, 2, 2, 5, 5, 8, 10, 10, 10, 15 >$) (notice this is same as as A's median, which might be used as an argument for A being the most reasonable).

For the variance in the aggregated sample we get a value ~ 1266.29. Overall aggregating the results produces an outcome which is more dispersed than (A) or (B).

e. With the new arrangment, we obtain an average,

$$\frac{10 + 2 \cdot 15 + 2 \cdot 20 + 2 \cdot 35 + 2 \cdot 45 + 2 \cdot 55 + 2 \cdot 64 + 3 \cdot 70 + 3 \cdot 85 + 2 \cdot 90}{21}$$

which equals 53.48. The median performance is now 55 and variance is found to be ~ 695.87.

There is an argument for viewing the replacement assessment as "fairer": average and median are similar (suggesting no bias to small numbers of very high or very low marks) and the variance is noticeably lower than the previous 4 cases (individual outcomes from A, B, and C, and aggregating all three earlier tests).

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).

Solutions

a. The best line fit can be shown to be $y = 0.19789x + 0.0276$. Recalling course textbook (p. 341):

$$W_x = \sum_{i=1}^{10} i = 55 \; ; \; W_y = \sum_{i=1}^{10} y_i = 11.16$$

$$W_{xy} \;=\; \sum_{i=1}^{10} i \cdot y_i \;\;=\;\; 77.706 \;\;;\;\; W_{xx} \;\;=\;\; \sum_{i=1}^{10} i^2 \;=\; 385$$

The gradient (m) of the best fit line is

$$\frac{10 \cdot W_{xy} - W_x W_y}{10 \cdot W_{xx} - (W_x)^2} \;\;=\;\; \frac{10 \cdot 77.706 - 55 \cdot 11.16}{10 \cdot 385 - (55)^2} \;\sim\; 0.19789$$

The offset (c) of this line being

$$\frac{W_x x W_y - W_{xy} W_x}{10 \cdot W_{xx} - (W_x)^2} \;\;=\;\; \frac{385 \cdot 11.16 \;-\; 77.706 \cdot 55}{10 \cdot 385 - (55)^2} \;=\; 0.0276$$

b. For Geometric regression we work with the (Natural) logarithms of the Data (x) and observation (y) sets (course textbook, p. 344) finding a best fit line for these data. The data we now work with are,

Table 3: Logarithm of Data/Observation Set

$\log x$	$\log y$
0	-1.661
0.693	-0.728
1.099	-0.446
1.386	-0.274
1.609	-0.051
1.792	0.316
1.946	0.285
2.079	0.419
2.197	0.701
2.303	0.642

By similar methods to those described in part (a) the best fit line for these data is

$$y \;\;=\;\; u_1 x + u_2 \;\;=\;\; 0.98422x - 1.56607$$

Recalling that we are interested not in a line but a **power function** via the approach described on pages 344–347 of the course text, the power function, $y = ax^b$ for the data of Table 2 is found to have $a = \exp(-1.56607)$ and $b = 0.98422$. In summary,

Best fit line for Table 2: $y = 0.19789x + 0.0276$

Best fit power function for Table 2: Since $\exp(-1.56607) \sim 0.2089$, $y = 0.2089x^{0.98422}$.

c,d The power function "appears" to fit these data more accurately. Computing the values

$$\sum_{i=1}^{10}(y_i - lin(x_i))^2 \quad ; \quad \sum_{i=1}^{10}(y_i - pow(x_i))^2$$

where $lin(x) = 0.98422x - 1.56607$ and $pow(x) = 0.2089x^{0.98422}$ the second quantity will be smaller.