Statistical Analysis for Engineers

Worksheet on Descriptive Statistics

Question 1

A set of Rosemary plants were grown in three different groups, differentiated by the mixes of soil, compost and lighting. Their heights were measured after a fixed period of time. The results are given (in cm) in the following table:

Heights (cms)	A	B	C
100 to 105	10	24	7
105 to 110	16	15	3
110 to 115	19	20	13
115 to 120	59	25	26
120 to 125	26	21	22
125 to 130	7	19	58
130 to 135	3	16	11

Set up a template on a spreadsheet or write an algorithm in the C programming language to carry out the following calculations:

1. Find the frequency mean and frequency standard deviation for each data set.
2. Calculate the median and quartiles for the two data sets.

Answer the remaining two questions:

3. Comment on the comparison between the means and standard deviations of the two datasets.

4. Comment on the comparison of the mean and median for each data set.

Question 2

The population mean and standard deviation, for a list of n numbers x_i, are defined by the expressions

$$\bar{x} = \frac{\sum x_i}{n}, \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n}.$$

Show that the standard deviation is also given by the following expressions:

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n}, \text{ or } s^2 = \frac{\sum x_i^2 - \left(\sum x_i\right)^2 / n}{n}.$$

Question 3

For a frequency distribution, the mean is given by the equation

$$\bar{x} = \frac{\sum_i m_i f_i}{n}.$$

This uses estimate $\sum_i m_i f_i$ for the sum. Let m_L be the lowest possible value of the mean for the frequency distribution and let m_H be the highest possible value of the mean. Identify what sums would be used to produce these two estimates and from this show that

$$\bar{x} = \frac{m_L + m_H}{2}.$$

Question 4

The frequency standard deviation is given by:

$$s^2 = \frac{\sum_i m_i^2 f_i - n\bar{x}^2}{n-1} = \frac{\sum_i (m_i - \bar{x})^2 f_i}{n-1}.$$

Show that these two versions of the equation are equal.

Question 4

The quantity $\sum_i m_i^2 f_i$ is the estimate for the sum of squares using the midpoints. Identify the lowest possible value of the sum of squares, call it S_L, and the highest, call it S_H. Then answer the following questions

1. Assuming that all groups have the same width w, show that

$$m_L = \bar{x} - \frac{w}{2} \text{ and } m_H = \bar{x} + \frac{w}{2}.$$

   [This is an alternative way to prove Question 3 above.]

2. Let s^2 be the usual estimate of the variance using the midpoints. Let s_L^2 be the estimate produced using the value S_L for the sum of the squares and m_L for the mean. Let s_U^2 the corresponding estimate using S_H and m_H. Show that $s^2 = s_L^2 = s_U^2$.
3. Addendum: show that, for a number p, if the midpoint m_i is replaced by values $L_i + pw$ for the estimate of the sum (and therefore the mean) and the sum of squares then the same value of s^2 is found.

Question 5

For a list of paired numbers (X_i, Y_i), the value of r is defined as

$$r = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2}\sqrt{\sum\left(Y_i - \overline{Y}\right)^2}}.$$

Answer the following questions.

1. Show that r is also given by

$$r = \frac{\sum X_i Y_i - \overline{X}\,\overline{Y}}{\sqrt{\sum X_i^2 - \overline{X}^2}\sqrt{\sum Y_i^2 - \overline{Y}^2}}.$$

2. Show that if the points are on a line, in other words, for every i in the list,

$$Y_i = mX_i + c,$$

for some m and c, then $r = +1$ or $r = -1$.

3. Show that if variable X is transformed to $a + bX$ and Y is transformed to $c + dY$ then the value of r does not change, provided both b and d are positive.

Question 6

A lecturer is seeking to prove that attendance is a strong predictor of final marks for a group of students. For each student, the attendance over the course of the semester and the final mark were recorded. The values are given in the table shown:

Attendance	45	32	67	56	78	86	43
Final Mark	56	34	62	76	65	74	33

1. Calculate the correlation coefficient r for this data.
2. Calculate the coefficients for the least squares linear equation that attempts to predict the final mark from the attendance.
3. Draw a scatterplot of the data points given above.

Question 7

For a list of paired numbers (X_i, Y_i), the value of r is defined as above. Treat the data and the equation we are trying to fit:

$$Y_i = \alpha + \beta X_i,$$

as a set of n equations for α and β, with the quantities X_i and Y_i as the coefficients. Finding α and β is solving the over-determined system

$$\alpha + \beta X_i = Y_i.$$

Denote the actual solution parameter vector by $(a, b)^{\mathrm{T}}$. Set X to be the matrix

$$X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{pmatrix}^T,$$

and Y to be the column vector with Y_i as the i-th element. Then to solve

$$X\begin{pmatrix} a \\ b \end{pmatrix} = Y,$$

we use the pseudo inverse

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left(X^T X\right)^{-1} X^T Y.$$

Show that this matrix equation gives the same equations for a and b as the calculus solution shown in your notes.