Find quartiles Q_1, Q_2, Q_3 and mode:

$x_1, \cdots, x_n$ — raw data

n — sample size

q_i — the rank (position) of Q_i

$[q_i]$ — the integer part of $q_i,\ i = 1, 2, 3.$ e.g. [3.3]=3, [4.8]=[4]

Find Q_2 (the second quartile or median):

Step 1: Take an ordered array: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Step 2: Find the position $q_2 = \frac{n+1}{2}$.

Step 3: Find Q_2.

Rule 1: if n is odd, $Q_2 = x_{(q_i)}$.

Rule 2: if n is even, $Q_2 = \frac{x_{([q_2])} + x_{([q_2]+1)}}{2}$.

Find Q_1 and Q_3 (the first and third quartiles):

Step 1: Step 1: Take an ordered array: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Step 2: Find the positions:

$$q_1 = \frac{n+1}{4}, \quad q_3 = \frac{3(n+1)}{4}.$$

Step 3: Find Q_1 and Q_3.

Rule 1: if q_i is an integer, $Q_i = x_{(q_i)},\ i = 1, 3.$

Rule 2: if q_i is a fractional half (e.g. 2.5, 4.5),

$$Q_i = \frac{x_{([q_i])} + x_{([q_i])+1}}{2}, \quad i = 1, 3.$$

Rule 3: if q_i is neither an integer nor a fractional half,

$$Q_i = x_{(\tilde{q}_i)}, \quad \tilde{q}_i \text{ is the nearest integer around } q_i.$$

Example 1:

$$39 \quad 29 \quad 43 \quad 52 \quad 39 \quad 44 \quad 40 \quad 31 \quad 44 \quad 35$$

a) Find median Q_2:

Step 1: 29 31 35 39 39 40 43 44 44 52
Step 2: Position $q_2 = \frac{n+1}{2} = \frac{11}{2} = 5.5 \quad (n = 10)$
Step 3: Median $Q_2 = \frac{x_{(5)}+x_{(6)}}{2} = \frac{39+40}{2} = 39.5.$

b) Find mode: there are two modes, 39 and 44.

c) Find Q_1 and Q_3:

$$q_1 = \frac{n+1}{4} = \frac{10+1}{4} = 2.75, \quad \tilde{q}_1 = 3, \quad Q_1 = x_{(3)} = 35,$$
$$q_3 = \frac{3(n+1)}{4}\frac{3 \times 11}{4} = 8.25, \quad \tilde{q}_3 = 8, \quad Q_3 = x_{(8)} = 44.$$

Example 2:

$$1 \quad 3 \quad 0 \quad 3 \quad 26 \quad 2 \quad 7 \quad 7 \quad 0 \quad 2 \quad 3 \quad 3 \quad 6 \quad 3$$

a) Find median Q_2:

Step 1: 0 0 1 2 2 3 3 3 3 3 4 6 7 26
Step 2: Position $q_2 = \frac{14+1}{2} = 7.5 \quad (n = 14)$
Step 3: Median $Q_2 = \frac{x_{(7)}+x_{(8)}}{2} = 3.$

b) Find mode: mode = 3

c) Find Q_1 and Q_3 :

$$q_1 = \frac{14+1}{4} = 3.75, \quad \tilde{q}_1 = 4, \quad Q_1 = x_{(4)} = 2,$$
$$q_3 = \frac{3(14+1)}{4} = 11.25, \quad \tilde{q}_3 = 11, \quad Q_3 = x_{(11)} = 4.$$

Example 3:

$$19.0 \quad 20.8 \quad 22.3 \quad 22.4 \quad 24.9 \quad 26.0 \quad 29.9$$

a) Find median Q_2:

Step 1: 19.0 20.8 22.3 22.4 24.9 26.0 29.9
Step 2: Position $q_2 = \frac{7+1}{2} = 4 \quad (n = 7)$
Step 3: Median $Q_2 = x_{(4)} = 22.4.$

b) Find mode: no mode

c) Find Q_1 and Q_3 :

$$q_1 = \frac{7+1}{4} = 2, \quad Q_1 = x_{(2)} = 20.8,$$
$$q_3 = \frac{3(7+1)}{4} = 6, \quad Q_3 = x_{(6)} = 26.0.$$

Moments for describing characteristics of data:

- Mean (central tendency):

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

- Variance (variation):

$$a_2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2$$

- Skewness (symmetry):

$$a_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$

$a_3 < 0$, left-skewed
$a_3 = 0$, symmetrical
$a_3 > 0$, right-skewed

- Kurtosis (peak):

$$a_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4}$$

$a_4 < 3$, less peaked than normal
$a_4 = 3$, same peaked as normal
$a_4 > 3$, more peaked than normal

E.g. $a_1 = 0, a_2 = 0.36, a_3 = -0.11, a_4 = 2.49$ indicates that

The mean is 0.

The standard deviation (std) is 0.6.

The data are slightly skewed to the left.

The data are not as peaked as normal.

Computation of some statistics

Computation of sample variance S^2:

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i^2 + \overline{x}^2 - 2\overline{x}x_i)}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}\sum_{i=1}^{n} x_i}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}{n-1} \quad (\because \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i) \end{aligned}$$

Computation of sample covariance $Cov(x, y)$:

$$\begin{aligned} Cov(x,y) &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n-1} \end{aligned}$$

Computation of correlation coefficient r:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{Cov(x,y)}{S_x S_y} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2}} \end{aligned}$$

More examples:

Example 4: Consider the following two samples:

Sample 1: 10, 9, 8, 7, 8, 6, 10, 6
Sample 2: 10, 6, 10, 6, 8, 10, 8, 6

(a) Calculate the sample range for both samples. Would you conclude that both samples exhibit the same variability? Explain.

(b) Calculate the sample standard deviations for both samples. Do these quantities indicate that both samples have the same variability? Explain.

(c) Write a short statement contrasting the sample range versus the sample standard deviation as a measure of variability.

Example 2: The concentration of a solution is measured six times by one operator using the same instrument. She obtains the following data:

63.2, 67.1, 65.8, 64.0, 65.1, and 65.3 (grams per liter).

(a) Calculate the sample mean and sample standard deviation.

(b) Suppose that the desirable value for this solution has been specified to be 65.0 grams per liter. Do you think that the sample mean value computed in (a) is close enough to the target value to accept the solution as conforming to target? Explain your reasoning.

(c) Suppose that in measuring the concentration, the operator must set up an apparatus and use a reagent material. What do you think the major sources of variability are in this experiment? Why is it desirable to have a small variance of these measurements?

Solutions:

1.

(a)

$$\begin{aligned} \text{Sample 1}: &\quad \text{Range} = 4 \\ \text{Sample 2}: &\quad \text{Range} = 4 \end{aligned}$$

Yes. The two appear to exhibit the same variability.

(b)

$$\begin{aligned} \text{Sample 1}: &\quad S = 1.604 \\ \text{Sample 2}: &\quad S = 1.852 \end{aligned}$$

No. Sample 2 has larger standard deviation.

(c) Range is a relatively crude measure of the sample variability as compared to S because S uses all data whereas range only uses two data points.

2.

(a) $\bar{X} = 65.083, \quad S = 1.367, \quad \text{CV} = (\frac{S}{\bar{X}}) \times 100\% = 2.1\%.$

(b) $\bar{X}$ is close enough to the target value to accept the solution as conforming due to small S and CV.

(c) (i) When the same setup is used for all measurements, a major source of variability might be in reagent material.

(ii) When each measurement uses a different setup, both reagent material and setup could be major sources of variability.

A low variance indicates the measurement error has low variability.