Variation in the data

Consider the following two datasets:

1st dataset: 49, 50, 51
2nd dataset: 0, 50, 100

Both datasets have the same mean: 50, but the 2nd dataset has more variability. We will discuss how to measure variability.

Measures

(a) Range
(b) Mean deviation from mean
(c) Variance and Standard deviation

Range

Range = largest value − smallest value.

Example

Data: 2, 4, 8, 5
Range = 8 – 2 = 6

- Range is not very useful. It only gives a rough idea about the variation.

Mean deviation from mean

$$\text{MD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

Note

For any data set

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Example

Data: 2, 3, 5, 6

$\bar{x} = 4$

$$\text{MD} = \frac{1}{4}(|2-4| + \cdots + |6-4|) = 1.5$$

Variance

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$$

Example

Data: 2, 3, 5, 6

$\bar{x} = 4$

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$$

$$= \frac{1}{4-1}((2-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2)$$

$$= 3.33$$

Note

The division is by $n-1$ because the number of free values (degrees of freedom) is $n-1$. If $n = 4$, and we know 3 values of $(x_i - \bar{x})$, the 4th one can be calculated.

Standard deviation (SD)

It is the positive square-root of variance and is denoted by s.

$s = \sqrt{s^2}$

Example

In the previous example, $s = \sqrt{3.33} = 1.83$

Note

Range, mean deviation, variance and SD cannot be negative.

Empirical Rule

If a distribution (histogram) appears to be symmetric and bell-shaped, we expect that approximately

- 68% of the data values will fall in the interval $(\bar{x} - s, \bar{x} + s)$
  (within one standard deviation of the sample mean)
- 95% of the data values will fall in the interval $(\bar{x} - 2s, \bar{x} + 2s)$
  (within two standard deviations of the sample mean)
- 99.7% of the data values will fall in the interval $(\bar{x} - 3s, \bar{x} + 3s)$
  (within three standard deviations of the sample mean)

Example

Let the mean and SD of commuting time (minutes) of workers be 60 and 10, respectively. Let the histogram be more or less symmetric and bell-shaped. We then have:

$\bar{x} - s = 60 - 10 = 50$
$\bar{x} + s = 60 + 10 = 70$

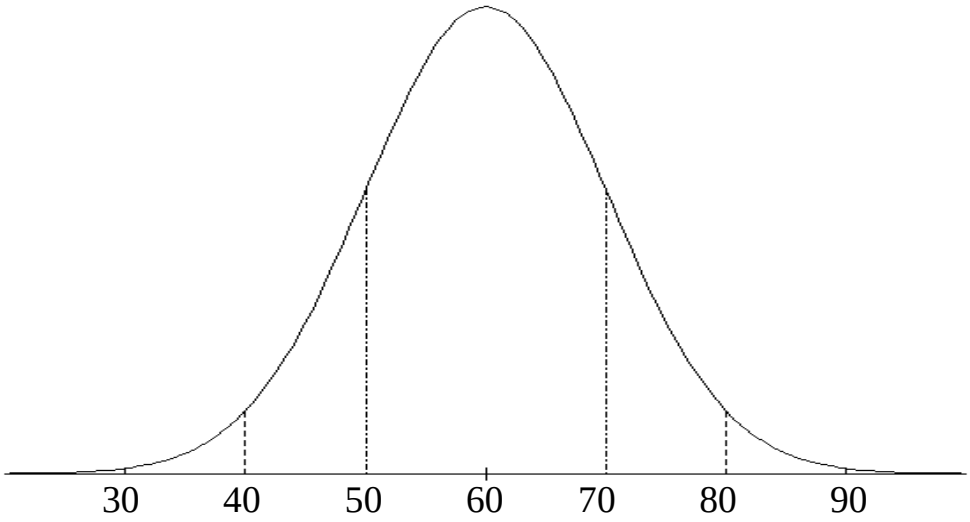
Approximately 68% workers have commuting time between 50 and 70 minutes.

$\bar{x} - 2s = 60 - 2 \times 10 = 40$
$\bar{x} + 2s = 60 + 2 \times 10 = 80$

Approximately 95% workers have commuting time between 40 and 80 minutes.

$\bar{x} - 3s = 60 - 3 \times 10 = 30$
$\bar{x} + 3s = 60 + 3 \times 10 = 90$

Approximately 97.7% workers have commuting time between 30 and 90 minutes.



Coefficient of variation (CV)

Let there be two datasets. First set contains small values and the second set contains large values. Comparing their standard deviations may be misleading. In order to compare their variability, we should use CV defined as

$$\text{CV} = \frac{s}{\bar{x}} \times 100\%$$

Example

The SD of a particular type of 10-mg tablets is 1 mg, while the SD of a particular type of 50-mg tablets is 2 mg. Which type of tablets has more variability?

Solution

For 10-mg tablets

$$\text{CV}(1) = \frac{s}{\bar{x}} \times 100\% = \frac{1}{10} \times 100\% = 10\%$$

For 50-mg tablets

$$\text{CV}(2) = \frac{s}{\bar{x}} \times 100\% = \frac{2}{50} \times 100\% = 4\%$$

Therefore, 10-mg tablets have more variability.