# Chapter IV

# Measures of Dispersion

**R. 4. 1.**
The measures of location indicate the general magnitude of the data and locate only the centre of a distribution. They do not establish the degree of variability or the spread out or scatter of the individual items and the deviation from the average.
Two distributions of statistical data may be symmetrical and have common arithmetic average, median and modes. Yet with these points in common they may differ widely in the scatter or in their values about the measures of location.

**D. 4. 1. (*Range*)**
1. The *range* of an *ungrouped* is defined by:

$$R := x_{\max} - x_{\min}$$

$x_{\min}$ : smallest value in the data set

$x_{\max}$ : largest value in the data set

2. The *range* of a *grouped* data is defined by:

$$R :\approx B_p - b_1$$

$b_1$ : lower bound of the first class

$B_p$ : upper bound of the class $p$ .

**Ex. 3. 1.** (*cont.*)
1. *Ungrouped*

$$R = 18.30 - 14.00 = 4.30 \ €$$

2. G*rouped*

$$R = 18.40 - 14.00 = 4.40 \ €$$

**R. 4. 2. (*Disadvantages of the Range*)**
1. It is influenced by outliers.
2. Its calculation is based on two values only: the largest and the smallest.

Thus, the range is not a very satisfactory measure of dispersion.

**D. 4. 2.** (***Interquartile Range***)
The *interquartile range* is defined as:

$$R_Q := x_{0.75} - x_{0.25}$$

**Ex. 3. 1.** (*cont.*)

1. *Ungrouped*

$$R_Q = 17.05 - 15.60 = 1.45 \text{ €}$$

2. G*rouped*

$$R_Q = 17.35 - 15.50 = 1.85 \text{ €}$$

**D. 4. 3.** (***Average Absolute or Linear Deviation - Ungrouped***)
1. *Absolute* (or *linear*) *average deviation from the median*:

$$d_{\tilde{x}_{0.5}} := \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}_{0.5}\right|$$

2. *Absolute* (or *linear*) *average deviation from the arithmetic average*:

$$d_{\bar{x}} := \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

**Ex. 3. 1.** (*cont.*)
1.

$$d_{\tilde{x}_{0.5}} := \frac{1}{30}\sum_{i=1}^{30}\left|x_i - 16.20\right| = \frac{1}{30}\cdot 27.8 \approx 0.93$$

2.

$$d_{\bar{x}} := \frac{1}{30}\sum_{i=1}^{30}\left|x_i - 16.30\right| = \frac{1}{30}\cdot 28.2 \approx 0.94$$

**R. 4. 3.** (***Minimum Property of the Median***)

$$\sum_{i=1}^{n}\left|x_i - \tilde{x}_{0.5}\right| \le \sum_{i=1}^{n}\left|x_i - N\right|, \qquad \forall\, N \in R^1$$

**D. 4. 4.** (***Average Absolute or Linear Deviation - Grouped***)

$$d_{\tilde{x}_{0.5}} := \approx \frac{1}{n}\sum_{i=1}^{p}\left|\bar{x}_i - \tilde{x}_{0.5}\right| \cdot F_i$$

$$d_{\bar{x}} := \frac{1}{n}\sum_{i=1}^{p}\left|\bar{x}_i - \bar{x}\right| \cdot F_i$$

**R. 4 . 4.**

In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoints*:

**Ex. 3. 1.** (*cont.*)
We had:

$$\tilde{x}_{0.5} = 16.29, \qquad \bar{x} = 16.40$$

| $i$ | $C_i$ | | $F_i$ | $m_i$ | $\left\|m_i - \tilde{x}_{0.5}\right\| \cdot F_i$ | $\left\|m_i - \bar{x}\right\| \cdot F_i$ |
|---|---|---|---|---|---|---|
| 1 | [14.00, | 15.00[ | 2 | 14.50 | 3.58 | 3.80 |
| 2 | [15.00, | 16.00[ | 11 | 15.50 | 8.69 | 9.90 |
| 3 | [16.00 | 17.00[ | 7 | 16.50 | 1.47 | 0.70 |
| 4 | [17.00 | 18.40[ | 10 | 17.70 | 14.10 | 13.00 |
| Total | | | 30 | | 27.84 | 27.40 |

$$d_{\tilde{x}_{0.5}} \approx \frac{27.84}{30} \approx 0.93, \qquad d_{\bar{x}} = \frac{27.4}{30} \approx 0.91$$

**D. 4. 5. (*Average Squared Deviation - Ungrouped*)**
1. A*verage squared deviation from the median*:

$$d^2_{\tilde{x}_{0.5}} := \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \tilde{x}_{0.5}\right)^2$$

2. *Average squared deviation from the arithmetic average (Variance)*

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \qquad \textit{(population variance)}$$

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \textit{(sample variance)}$$

3. *Standard deviation*

$$\sigma := \sqrt{\sigma^2}, \quad \sigma > 0 \qquad \textit{(population standard deviation)}$$

$$s := \sqrt{s^2}, \quad s > 0 \qquad \textit{(population standard deviation)}$$

**Ex. 3. 1.** (*cont.*)
1.

$$d^2_{\tilde{x}_{0.5}} := \frac{1}{30}\sum_{i=1}^{30}\left(x_i - 16.20\right)^2 = \frac{1}{30}\cdot 37.79 \approx 1.26$$

2.

$$s^2 := \frac{1}{29}\sum_{i=1}^{30}\left(x_i - 16.30\right)^2 = \frac{1}{29}\cdot 37.49 \approx 1.30$$

3.

$$s := \sqrt{s^2} \approx 1.14$$

**R. 4. 4.** (***Short-Cut Formulas for the Variance for Ungrouped Data***)

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{N}}{N}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

**D. 4. 6.** (***Average Squared Deviation - Grouped***)

$$d^2_{\tilde{x}_{0.5}} := \approx \frac{1}{n}\sum_{i=1}^{p}\left(\bar{x}_i - \tilde{x}_{0.5}\right)^2 \cdot F_i$$

$$\sigma^2 := \approx \frac{1}{N}\sum_{i=1}^{p}\left(\bar{x}_i - \mu\right)^2 \cdot F_i \qquad \textit{(population variance)}$$

$$s^2 := \approx \frac{1}{n-1}\sum_{i=1}^{p}\left(\bar{x}_i - \bar{x}\right)^2 \cdot F_i \qquad \textit{(sample variance)}$$

**R. 4 . 5.**

In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoint*.:

**Ex. 3. 1.** (*cont.*)
We had:

$$\tilde{x}_{0.5} = 16.29, \qquad \bar{x} = 16.40$$

| $i$ | $C_i$ | $F_i$ | $m_i$ | $\left(m_i - \tilde{x}_{0.5}\right)^2 \cdot F_i$ | $\left(m_i - \bar{x}\right)^2 \cdot F_i$ |
|---|---|---|---|---|---|
| 1 | [14.00, 15.00[ | 2 | 14.50 | 6.4082 | 7.22 |
| 2 | [15.00, 16.00[ | 11 | 15.50 | 6.8651 | 8.91 |
| 3 | [16.00 17.00[ | 7 | 16.50 | 0.3087 | 0.07 |
| 4 | [17.00 18.40[ | 10 | 17.70 | 19.8810 | 16.90 |
| Total | | 30 | | 33.4630 | 33.10 |

$$d^2_{\tilde{x}_{0.5}} \approx \frac{33.4630}{30} \approx 1.12, \qquad s^2 = \frac{33.10}{29} \approx 1.14, \qquad s \approx 1.07$$

**R. 4. 6.** (***Short-Cut Formulas for the Variance for Grouped Data***)

$$\sigma^2 = \frac{\sum_{i=1}^{n} \bar{x}_i^2 \cdot F_i - \frac{\left(\sum_{i=1}^{n} \bar{x}_i \cdot F_i\right)^2}{N}}{N} \qquad \textit{(population variance)}$$

$$s^2 = \frac{\sum_{i=1}^{n} \bar{x}_i^2 \cdot F_i - \frac{\left(\sum_{i=1}^{n} \bar{x}_i \cdot F_i\right)^2}{n}}{n-1} \qquad \textit{(sample variance)}$$

**T. 4. 1.** (***Chebyshev***)

For any number $k > 1$, at least $\left(1 - 1/k^2\right)$ of the data values lie within $k$ standard deviations of the mean.

**Ex. 4. 1.**

The arithmetic mean biweekly amount contributed by the employees of a company to company's profit-sharing plan is \$51.54, and the standard deviation is \$7.51.
At least what percent of the contributions lie within plus 3.5 standard deviations and minus standard deviations of the mean?

*Solution:*

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 0.92.$$

**R. 4. 7.** (***Empirical Rule, Three-Sigma Rule***)
Whereas Chebyshev's theorem is applicable to any kind of distribution, the *empirical rule* applies only to a specific type of distribution called a *normal distribution*:

For a normal distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean
2. 95% of the observations lie within two standard deviations of the mean
3. 99.7% of the observations lie within three standard deviations of the mean

**R. 4. 8.**
*At least* 50% of the observations in a data set lie in the interval

$$[\mu-\sigma,\quad \mu+\sigma], \qquad \textit{(population)}$$

$$\left[\bar{x}-s,\quad \bar{x}+s\right] \qquad \textit{(sample)}$$

**D. 4. 7.** (***Coefficient of Variation***)

$$v := \frac{\sigma}{\mu} \qquad \textit{(population)}$$

$$v := \frac{s}{\bar{x}} \qquad \textit{(sample)}$$

**R. 4 . 9.**
The coefficient of variation is used to compare the variability of data sets with different arithmetic averages.

**Ex. 3. 1.** (*cont.*)
1. *Ungrouped:*

$$v := \frac{1.14}{16.30} \approx 0.07$$

2. *Grouped:*

$$v := \frac{1.07}{16.40} \approx 0.07$$

**R. 4 . 7** (***A Rule of Thumb***)
The arithmetic average of a data set is to be considered as representative only if its coefficient of variation is less than 0.5 (or 50%).

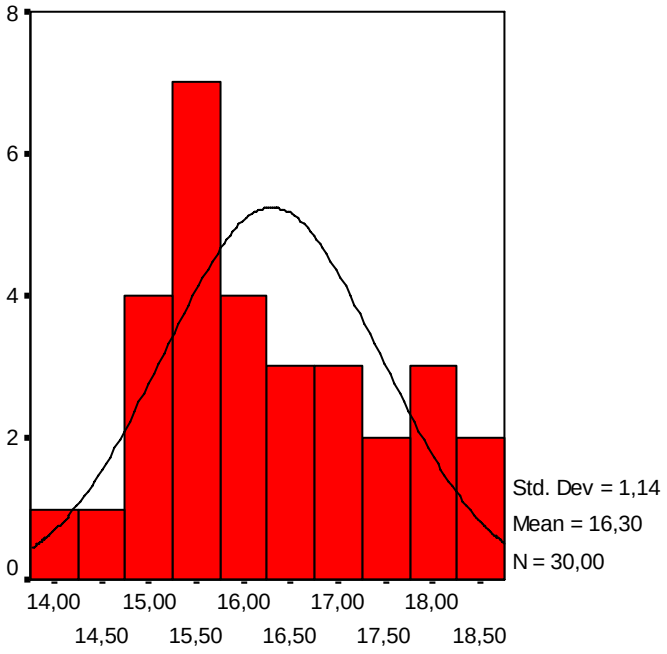
**D. 4. 7.** (***Skewness***)
The *coefficient of skewness* is defined as follows:

$$S := \frac{3\left(\bar{x} - Me\right)}{s}, \qquad (-3 \le S \le +3)$$

**Ex. 3. 1.**

1. *Ungrouped*

$$S = \frac{3(16.30 - 16.20)}{1.12} \approx 0.27$$




2. *Grouped*

$$S \approx \frac{3(16.40 - 16.29)}{1.12} \approx 0.29$$

**D. 4. 8.** **(*Box-and-Whisker Plot*)**
A *box-and-whisker Plot* is a plot that shows the centre, spread, and skewness of a data set.
**R. 4. 10.**
It is constructed by drawing a box and two whiskers that use the median, the first (lower) quartile, the third (upper) quartile, and the smallest and the largest values in the data set between the lower and the upper inner *fences*.
The following example explains all the steps needed to make a box-and-whisker plot

**Ex. 4. 2.**
The following data are the incomes [in thousands of €] for a sample of 12 households:

35 29 44 72 34 64 41 50 54 104 39 58

Construct a box-and-whisker plot of these data.

*Solution:*

*Step 1*
Find the values of the median, the first quartile, the third quartile, and the interquartile range:

$$x_{0.5} = 47, \quad x_{0.25} = 37, \quad x_{0.75} = 61, \quad R_Q = 61 - 37 = 24.$$

*Step 2*
Find the points that are $1.5 \cdot R_Q$ below $x_{0.25}$ and $1.5 \cdot R_Q$ above $x_{0.75}$. These two points are called the *lower* and the *upper inner fences*, respectively:

$$1.5 \cdot R_Q = 1.5 \cdot 24 = 36$$

Lower inner fence = $x_{0.25} - 36 = 37 - 36 = 1$

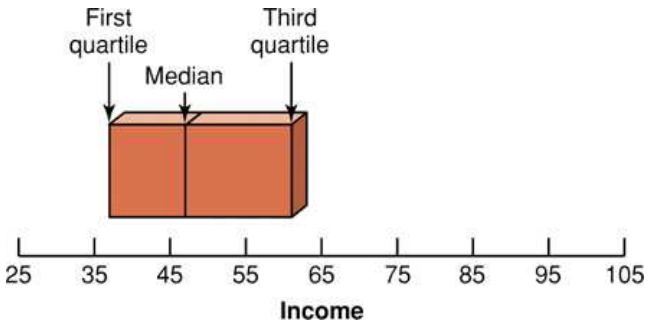
Upper inner fence = $x_{0.75} + 36 = 61 + 36 = 97$.

*Step 3*
Determine the smallest and the largest values in the given data set within the two inner fences:

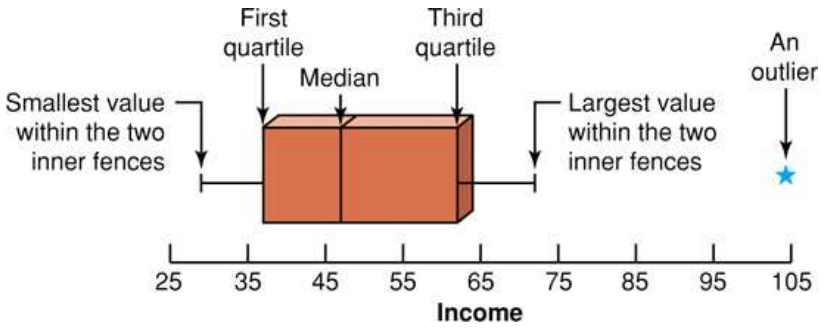
Smallest value within the two inner fences = 29

Largest value within the two inner fences = 72

*Step 4*
Draw a horizontal line and mark the income levels on it such that all the values in the given data set are covered. Above the horizontal line, draw a bow with its left side at the position of the first quartile and the right side at the position of the third quartile. Inside the box, draw a vertical line at the position of the median: The result is shown in the following figure:




*Step 5*
By drawing two lines, join the points of the smallest and the largest values within the two inner fences to the box. These values are 29 and 72 in this example. The two lines that join the box to these two values are called *whiskers*. A value that falls outside the two inner fences is shown by making an asterisk and is called an outlier. This completes the box-and-whisker plot, as shown in the following figure:




In the above figure, about 50% of the data values fall within the box, about 25% of the values fall on the left side of the box, and about 25% fall on the right side of the box. Also, 50% of the values fall on the left side of the median and 50% lie on the right side of the median. The data of this example are skewed to the right because the lower 50% of the values are spread over a smaller range than the upper 50% of the values.

**R. 4. 11.**

The observations that fall outside the two inner fences are called *outliers*. These outliers can be classified into two kinds: mild and extreme outliers.

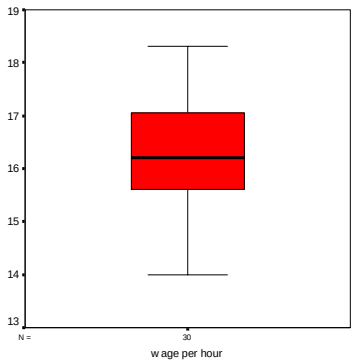
To do so we define two outer fences – a lower *outer fence* at $3.0 \cdot R_Q$ before the first quartile and an *upper outer fence* at $3.0 \cdot R_Q$ above the third quartile. If an observation is outside either of the two inner fences but within either of the two outer fences, it is called a *mild outlier*. An observation that is outside either of the two outer fences, is called an *extreme outlier*.

For the above example, the outer fences are -35 and 133. Because 144 is outside the upper inner fence but inside the upper outer fence, it is a mild outlier.

For a symmetric data set, the line representing the median will be in the middle of the box and the spread of the value will be over almost the same range on both side of the box.

**Ex. 3. 1.** (*cont.*)

Box-and-Whisker Plot:



*(Last revised: 31.04.09)*