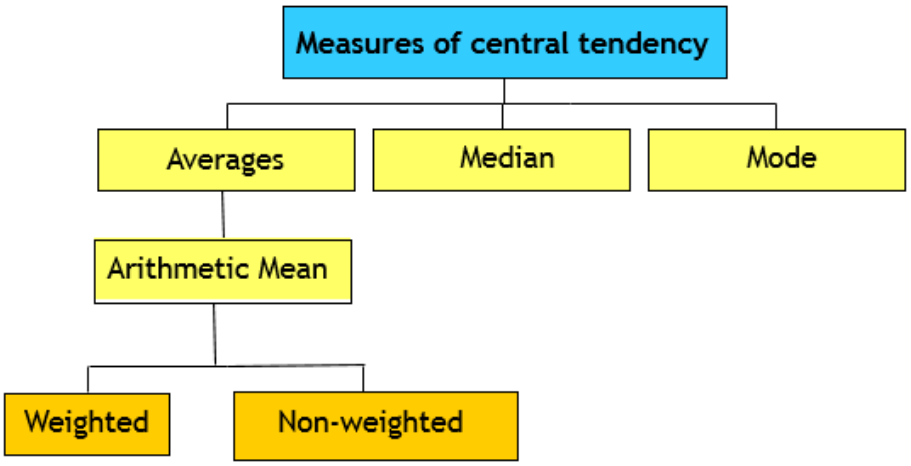
UNIT 4. NUMERICAL MEASURES: LOCATION MEASURES

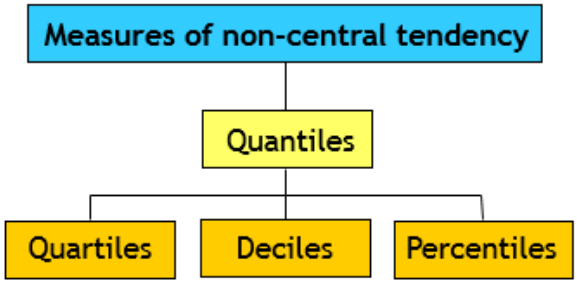
Location measures try to situate where the frequency distribution is located, looking for the most representative values (centre) or the extreme values (tails). They are used to study the characteristics of values that indicate the position of a group of values. Their purpose is to describe and synthesise the information contained in a set of data, usually for the purpose of comparison with other data.

They are classified into two groups: (i) Measures of central tendency, which try to situate central or representative values of the distribution; and (ii) Measures of non-central tendency, which try to situate intermediate and extreme values of the distribution.

We will consider the following measures of central tendency[1]:



And the following measures of non-central tendency:



[1] Within the averages, we can also find the geometric mean and the harmonic mean.

MEASURES OF CENTRAL TENDENCY

A measure of central tendency is a value that attempts to describe the data set of a variable by identifying the central position within the data set. The most commonly used measures of central tendency are: the average or arithmetic mean (weighted or unweighted), the median and the mode.

Arithmetic mean or average

For a frequency distribution of a set of data $\{(x_i, n_i); i=1,...,k\}$ with $N = n_1+n_2+...+n_k$ being the number of observations in the data set, we define the **arithmetic mean** or **average** as the ratio of the sum of all the values of the variable to the total number of subjects in the population:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{k} x_i \times n_i = \sum_{i=1}^{k} x_i \times f_i$$

If data are grouped, we will use class midpoints as the values of x_i. It can not be calculated for qualitative variables except for binary variables codified with 0/1 (in this case the arithmetic mean is a proportion).

The average fulfils the following properties:

- It is unique.
- All the values are included in computing the arithmetic mean.
- It only makes sense for quantitative variables and it is employed preferably in continuous variables.
- It can not be calculated for grouped variables with some unlimited classes.
- It is not a robust measure (it is affected by unusually large or small values).
- It can be considered as a balance point for the set of data (the sum of the deviations of each value from the mean is zero). Formally:

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

- It minimises the mean of squared deviations. Formally:

$$Min\ \ e(x) = \frac{\sum_i (x - x_i)^2 \times n_i}{N} \rightarrow x_{opt} = \frac{\sum_i x_i \times n_i}{N} = \bar{x}$$

- It is not invariant for linear changes (scale and origin changes). Formally, it is satisfied that if $Y = a + b \cdot X$, then $\bar{y} = a + b \cdot \bar{x}$
- It is a separable measure. Formally, for two subsets of sizes N_A and N_B, with associated averages $\bar{x}_A$ and $\bar{x}_B$, the total average can be obtained as:

$$\bar{x}_T = \frac{N_A \times \bar{x}_A + N_B \times \bar{x}_B}{N_A + N_B}$$

Weighted means are employed when some data points count more strongly than others. Weights (w_i) indicate the importance of each value. The weighted arithmetic mean is calculated as:

$$\overline{x}_w = \frac{\sum_{i=1}^{k} x_i \times w_i}{\sum_{i=1}^{k} w_i}$$

where $\{w_i; i = 1, \ldots, k\}$ is a set of weights assigning the importance of each piece of information.

Example

What is the average number of hours of study per day for this group of students?

Number of hours of study per day	Number of students	
x_i	n_i	f_i
1	5	10%
2	15	30%
3	20	40%
4	8	16%
5	2	4%
	50	

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \times n_i}{N} = \frac{1 \times 5 + 2 \times 15 + 3 \times 20 + 4 \times 8 + 5 \times 2}{50} = \frac{137}{50} = 2.74 \text{ hours}$$

$$\bar{x} = 1 \times 0.10 + 2 \times 0.30 + 3 \times 0.40 + 4 \times 0.16 + 5 \times 0.04 = 2.74 \text{ hours}$$

On average, students study 2.74 hours per day on average.

Example

The following table shows the frequency distribution of the monthly salary of the workers of a company:

Monthly salary (€)	Number of workers
500-1000	50
1000-1500	150
1500-2000	200
2000-2500	90
2500-3000	10

What is the average monthly salary of the workers of this company?

X- Monthly salary (€)		n_i	x_i	$n_i x_i$
500	1000	50	750	37,500
1000	1500	150	1250	187,500
1500	2000	200	1750	350,000
2000	2500	90	2250	202,500
2500	3000	10	2750	27,500
	Total	500		805,000

$$\bar{X} = \frac{805{,}000}{500} = 1{,}610€$$

The average monthly salary is 1,610 € (This is an *approximate* value).

Mode

For a frequency distribution of a variable $\{(x_i, n_i); i=1,...,k\}$ with $N=n_1+n_2+...+n_k$ being the number of observations in the data set, we define the **MODE** (Mo) as the most frequent value in the data set.

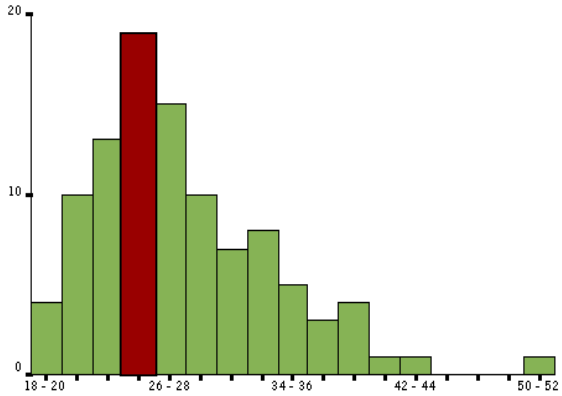
The calculation depends on the type of variable:

- For **qualitative or quantitative discrete** data:

$$Mo = x_j \quad \text{for which} \quad n_j = \max_i\{n_i\}$$

- For **continuous or grouped data**: it is necessary to follow this sequence:
  (1) Calculate the frequency densities: $d_i = n_i/a_i$ or $d_i = f_i/a_i$
  (2) Identify the "Modal class" $(L_{m-1}, L_m]$ which is the one with the greatest frequency **density.**
  (3) Approximate the mode (*Mo*) using the expression:

$$Mo = L_{m-1} + \frac{\mathrm{d}_{m+1}}{d_{m-1} + d_{m+1}} \cdot a_m$$



Or, in a simplified version, to keep the class midpoint of the modal class:

$$Mo = x_m = \frac{L_{m-1} + L_m}{2}$$

The mode fulfils the following properties:

- It does not have to be unique.
- If there are several modes, this usually means that the population does not show homogeneous behaviour with respect to the variable and, therefore, several groups exist (see following example).
- It does not employ all the values from the frequency distribution.
- It can be calculated for any type of variable (all measurement scales).
- It is more robust than the arithmetic mean (not affected by extremely high or low values).
- It is not invariant for linear changes: If Y = a + bX, then: Mo(Y) = a + b·Mo(X)
- It is not a separable measure.
- It is easy to interpret and simple to calculate.

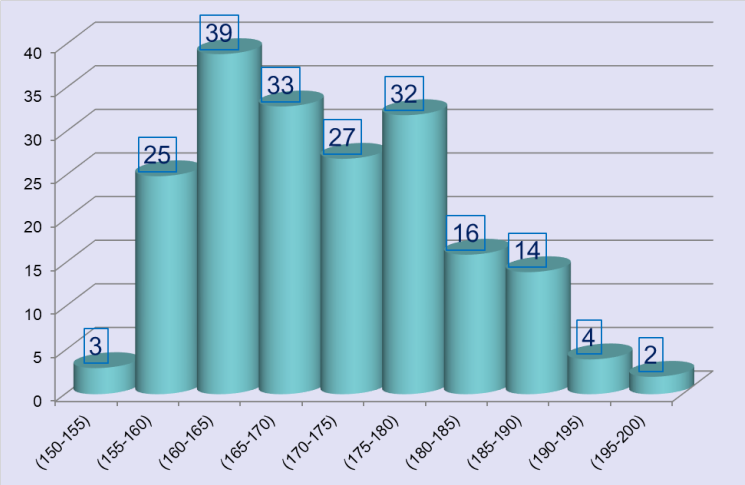
Example

Depending on the number of modes, the distribution is classified as **unimodal, bimodal or multimodal**. If it is not unique, the mode loses representativeness. The existence of two or more modes is usually due to a mixture of two or more heterogeneous groups so that it is convenient to study the groups separately. In the

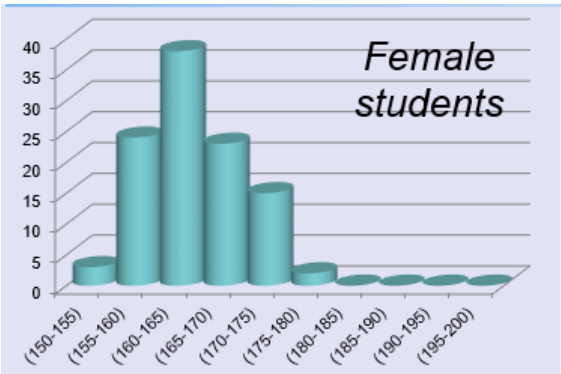
following example, we observe a bimodal distribution corresponding to the heights of a set of students:

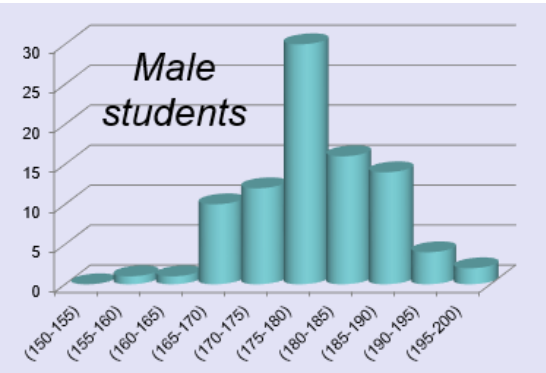
Height of students

(150-155)	3
(155-160)	25
(160-165)	39
(165-170)	33
(170-175)	27
(175-180)	32
(180-185)	16
(185-190)	14
(190-195)	4
(195-200)	2



Conducting the study separately:





Example

What is the most frequent number of hours of study per day for this group of students?

Number of hours of study per day	Number of students	
x_i	n_i	f_i
1	5	10%
2	15	30%
3	20	40%
4	8	16%
5	2	4%
	50	

The most frequent number of hours of study per day in this group of students is 3 hours.

Example

The following table shows the frequency distribution of the monthly salary of the workers of a company:

Monthly salary (€)	Number of workers
500-1000	50
1000-1500	150
1500-2000	200
2000-2500	90
2500-3000	10

What is the most frequent monthly salary of the workers of this company?

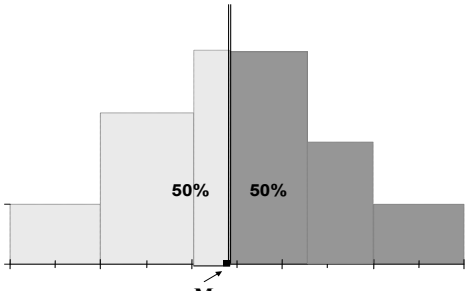
Li-1	*Li*	*ni*	*ai*	*di*
500	1000	50	500	0.10
1000	1500	150	500	0.30
1500	2000	200	500	0.40
2000	2500	90	500	0.18
2500	3000	10	500	0.02
		500		

$$Mo = \frac{L_{m-1} + L_m}{2} = \frac{1{,}500 + 2{,}000}{2} = 1{,}750€$$

The most frequent salary in this company is approximately 1,750 €.

Median

For a frequency distribution of an ordinal or quantitative variable: $\{(x_i, n_i); i=1,...,k\}$ with $N = n_1 + n_2 + ... + n_k$ being the number of observations in the data set, we define the **median** (Me) as the value that divides the distribution into two halves, leaving 50% below and 50% above.



In order to calculate it we must take into account the type of variable.

- For **qualitative ordinal** or **quantitative discrete** data (**non-grouped**):

  We look for the value whose cumulative absolute frequency is N/2. First, arrange all the observations from lowest value to highest value and find the integer number m so that m-1< N/2 ≤ m. Then:

  - If N is an odd number, the median is: Me = x_m
  - If N is an even number and the variable is quantitative, the median is usually defined as the average of the two middle values:

$$Me = \frac{x_m + x_{m+1}}{2}$$

Example

a) Calculate the median for the following set of data: 1, 3, 4, 5, 6, 7, 9

Me = 5 (N = 7 -> N/2 = 3.5 -> m = 4)

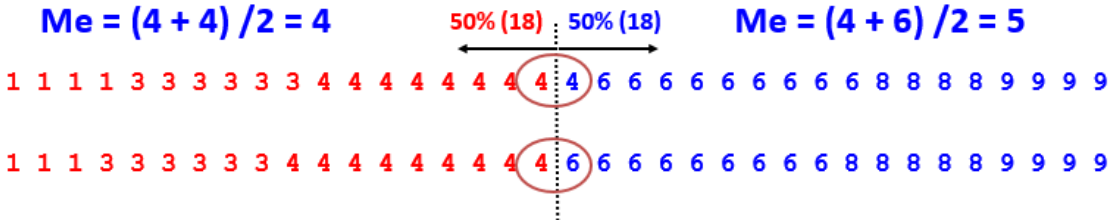
b) Calculate the median for the following set of data: 1, 3, 5, 6, 7, 9

Me = (5 + 6) /2 = 5.5 (N = 6 -> N/2 = 3 -> *m* = 3)

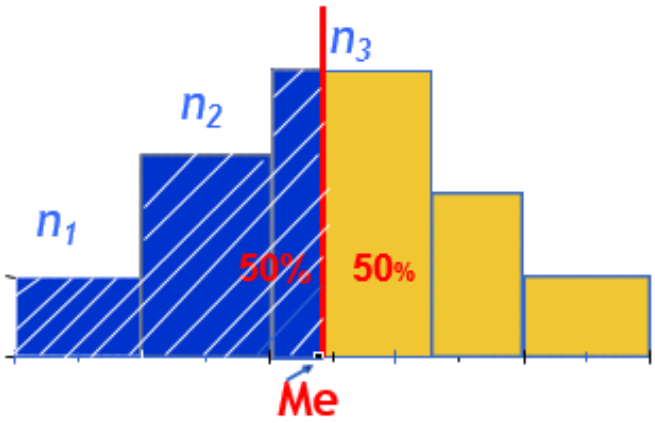
c) Calculate the median for the following sets of data:

X_i	n_i	N_i	F_i
1	4	4	11%
3	6	10	28%
4	9	19	53%
6	9	28	78%
8	4	32	89%
9	4	36	100%

X_i	n_i	N_i	F_i
1	3	3	8%
3	6	9	25%
4	9	18	50%
6	9	27	75%
8	5	32	89%
9	4	36	100%



- For **quantitative continuous** or **grouped data**:
  - Calculate the cumulative frequencies (N_i or F_i).
  - Identify the "**median class**" $(L_{m-1}, L_m]$ which is the one with a cumulative frequency greater than 50%: $N_{m-1} < N/2 \leq N_m$ or $F_{m-1} < 50\% \leq F_m$.
  - Approximate the median (Me) using the following rule:
    - If $F_m = 50\%$: $\boldsymbol{Me = L_m}$
    - If $F_m > 50\%$: $\boldsymbol{Me = x_m = \frac{L_{m-1}+L_m}{2}}$



The specific value can also be determined within this range by linear interpolation:

$$Me = L_{m-1} + \frac{N/2 - N_{m-1}}{n_m} \cdot a_m \;\; o \;\; Me = L_{m-1} + \frac{0.5 - F_{m-1}}{f_m} \cdot a_m$$

The median fulfils the following properties:

- It is unique, although it can be approximated in different ways (such as the class midpoint of the median class, or on the basis of proportionality/interpolation).
- It does not employ all the values from the frequency distribution.
- It does not make sense for nominal variables.
- It minimises the mean of absolute deviations.
- It is not invariant for linear changes (scale and origin changes). Formally, if $Y = a + b \cdot X$, then $Me(Y) = a + b \cdot Me(X)$.
- It is more robust than the arithmetic mean (it is not affected by extremely large or small values).
- It is not a separable measure.

Example

The following table shows the frequency distribution of the monthly salary of the workers of a company:

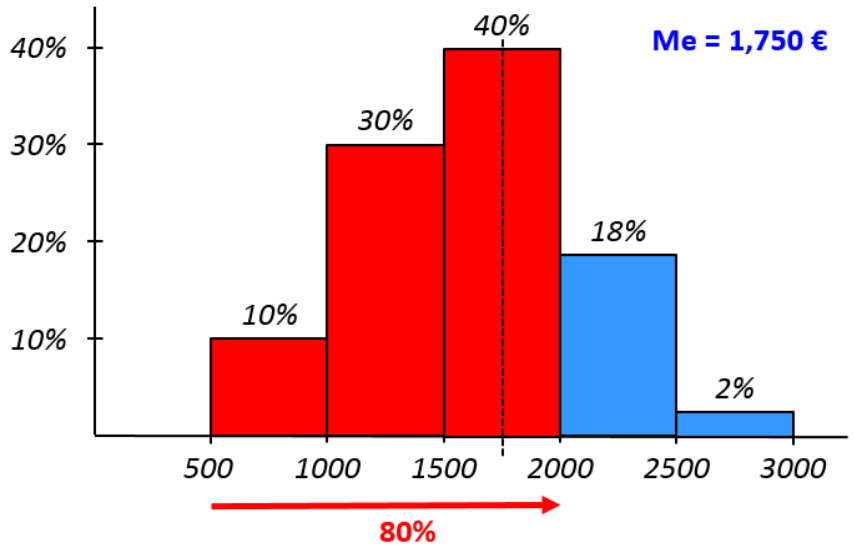
Monthly salary (€)	Number of workers
500-1000	50
1000-1500	150
1500-2000	200
2000-2500	90
2500-3000	10

What is the monthly salary that 50% of these workers earn more than?

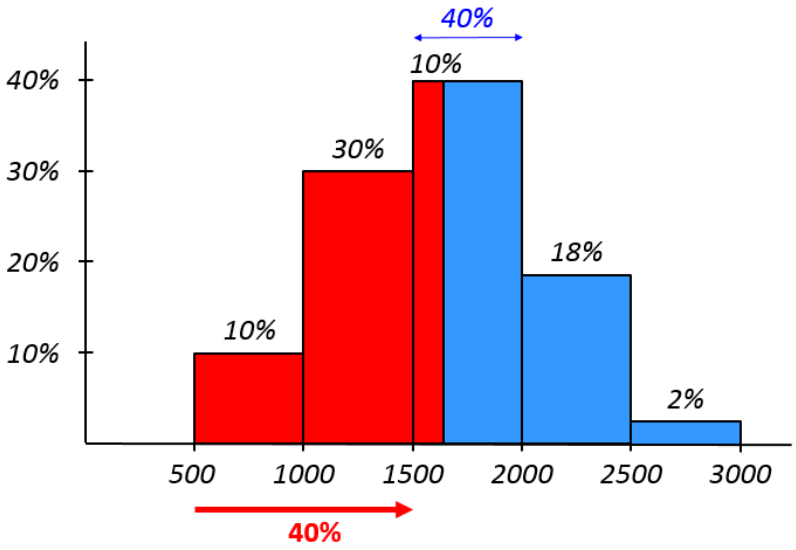
Li-1	Li	ni	fi	ai	di	Ni	Fi
500	1000	50	10%	500	0.1000	50	10.00%
1000	1500	150	30%	500	0.3000	200	40.00%
1500	2000	200	40%	500	0.4000	400	80.00%
2000	2500	90	18%	500	0.1800	490	98.00%
2500	3000	10	2%	500	0.0200	500	100.00%
		500					

F_3 = 80% > 50%

Then, median class is the 3rd one: **(1,500; 2,000] Then, Me = (1,500 + 2,000)/2 = 1,750 €**



If we **approximate** the median (*Me*) by interpolation, then: $Me = 1500 + \frac{50\%-40\%}{40\%} \times 500 = 1625$ €

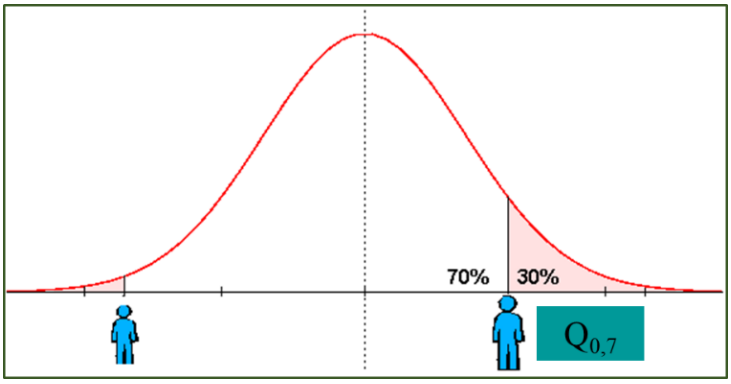


Measures of non-central tendency

Measures of non-central tendency, called **quantiles**, are measures that will not reflect any central tendency of the data set. Extending the concept of the median, they divide the frequency distribution into several parts, all of which have the same frequency, i.e. they divide the distribution into several ranges, all of which contain the same number of data. In this way, they reflect the upper, middle and lower values.

We define the ***p* quantile (Q_p)** as the **value** that **divides** the **distribution** into **two parts**, leaving the proportion ***p*·100% below** and (1-*p*)·100% above.

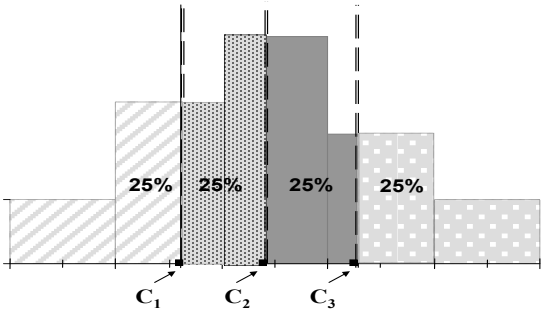
It is an **extension** of the **concept** of **median** changing the frequency 0.5 (50%) to any other value of *p*, with $0 < p < 1$ ($0 < p \cdot 100\% < 100\%$).



The most important quantiles are the **quartiles**, the **deciles** and the **percentiles**.

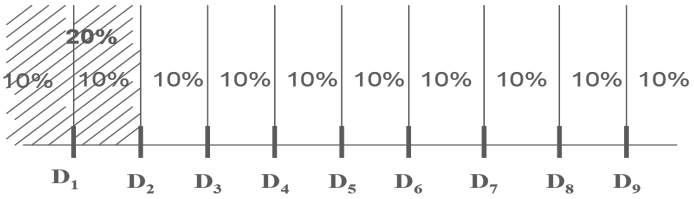
Quartiles

They divide the distribution into four equal parts. They are denoted by {Ci; i=1,2,3} and correspond to the 0.25; 0.50 and 0.75 quantiles: $C_1=Q_{0.25}$; $C_2=Q_{0.5}=Me$; $C_3=Q_{0.75}$



Deciles

They divide the distribution into ten equal parts. They are denoted by {Di; i = 1, 2, …,9} and correspond to the 0.1; 0.2;…; and 0.9 quantiles. Di = Qi/10 with i = 1, 2, …, 9.



Percentiles

They divide the distribution into a hundred equal parts. They are denoted by {Pi; i = 1, 2, …, 99} and correspond to the 0.01; 0.02;…; and 0.99 quantiles. $P_i = Q_{i/100}$ with i = 1,2, …, 99.

Calculation of the p quantile (Qp)

Its identification or calculation, which differs depending on the type of data, is essentially similar to the case of the median:

- First, calculate the cumulative frequencies N_i o F_i
- Find $\boldsymbol{m}$ that $\boldsymbol{F_{m-1} < 100 \times p\% \leq F_m}$ or $\boldsymbol{N_{m-1} < N \times p \leq N_m}$

- If data have not been grouped into classes
  - If $\boldsymbol{F_m > 100 \times p\%}$: $\boldsymbol{Q_p = x_m}$
  - If $\boldsymbol{F_m = 100 \times p\%}$: $\boldsymbol{Q_p = \frac{x_m + x_{m+1}}{2}}$
- If data have been grouped into classes:
  - If $F_m > 100 \times p\%$: Q_p belongs to the class (L_{m-1}, L_m) representant: $Q_p = \frac{L_{m-1} + L_m}{2}$
  - If $F_m = 100 \times p\%$: $Q_p = L_m$

The quantiles fulfil the following properties:

- They are unique.
- Their robustness depends on the value of p: if near to 0 or 1 they are not very robust; if near to 0.5 they are very robust.
- They are meaningless for nominal variables.
- They are not invariant for linear changes: If Y = a + b·X, then: Qp(Y) = a + b·Qp(X)
- They are not separable measures.

Example

Given the following frequency distribution, calculate the quartiles:

X_i	n_i	N_i	F_i
1	4	4	11.11%
3	6	10	27.77%
4	9	19	52.77%
6	9	28	77.77%
8	4	32	88.89%
9	4	36	100.00%

$F_2 = 27.77\% > 25\%$ → **$C_1 = 3$**

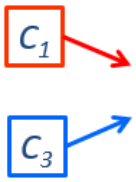
$F_4 = 77.77\% > 75\%$ → **$C_3 = 6$**

Example

The following table shows the frequency distribution of the monthly salary of the workers of a company:

Monthly salary (€)	Number of workers
500-1000	50
1000-1500	150
1500-2000	200
2000-2500	90
2500-3000	10

a) What are the salaries that limit the middle 50% of the workers?



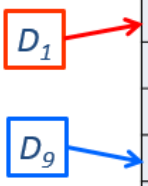
Li-1	Li	ni	fi	ai	Ni	Fi
500	1000	50	10%	500	50	10%
1000	1500	150	30%	500	200	40%
1500	2000	200	40%	500	400	80%
2000	2500	90	18%	500	490	98%
2500	3000	10	2%	500	500	100%
		500				

C_1 in (1,000-1,500] → $C_1 = 1{,}250$ €

C_3 in (1,500-2,000] → $C_3 = 1{,}750$ €

The middle 50% of the workers earn between 1,250 and 1,750 €.

b) What are the salaries that limit the middle 80% of the workers?



Li-1	Li	ni	fi	ai	Ni	Fi
500	1000	50	10%	500	50	10%
1000	1500	150	30%	500	200	40%
1500	2000	200	40%	500	400	80%
2000	2500	90	18%	500	490	98%
2500	3000	10	2%	500	500	100%
		500				

D_1 in (500-1,000] → $D_1 = 1{,}000$€

D_9 in (2,000-2,500] → $D_9 = 2{,}250$ €

The middle 80% of the workers earn between 1,000 and 2,250 €.

c) What is the salary that only 1% of the workers earn more than?

Li-1	*Li*	*ni*	*fi*	*ai*	*Ni*	*Fi*
500	1000	50	10%	500	50	10%
1000	1500	150	30%	500	200	40%
1500	2000	200	40%	500	400	80%
2000	2500	90	18%	500	490	98%
2500	3000	10	2%	500	500	100%
		500				

P_{99}

$P_{99} \in (2{,}500\text{-}3{,}000] \rightarrow P_{99} = 2{,}750$ € Only 1% of the workers earn more than 2,750 €.