Chapter 1. Descriptive Statistics: Position Parameters

Statistical parameters

Statistical parameters

are numbers that are intended to summarize the essential information relating to a quantitative statistical variable.

The **position parameters** indicate the “typical” value around which the observations are distributed.
The two most important position parameters are the **mean** and the **median.**

Fractiles , including **quartiles** and **deciles,** provide more detailed information about the series. They are a generalization of the median.
Dispersion parameters measure how much the observations deviate from the central value. The most important is **the standard deviation.**

Average

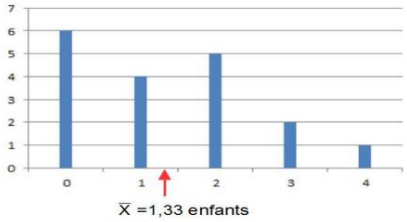
The **mean** of a variable is the sum of the observed values Xi divided by the total number n of individuals in the population:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

Example. On the right, the raw series from a survey on the number children from 18 families in a building.

$\overline{x} = \frac{0+0+0+0+0+0+1+1+1+1+2+2+2+2+2+3+3+4}{18} \simeq 1,33$

There are on average 1.33 children per family.



Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	4

Calculation of the average from the numbers and frequencies

The mean can be calculated from modalities y1, . . . , yp and of the associated ni numbers or f1 frequencies .

$$\overline{x} = \frac{y_1 \times n_1 + \cdots + y_p \times n_p}{n} = y_1 \times f_1 + \cdots + y_p \times f_p$$

Example. From the staffing table

No. of children yi	0	1	2	3	4	Total
Effective ni	6	4	5	2	1	18

$\overline{x} = \frac{0\times6+1\times4+2\times5+3\times2+4\times1}{18} \simeq 1,33$

From the frequency table:

No. of children yi	0	1	2	3	4	Total
Frequency fi	33%	22%	28%	11%	6%	100%

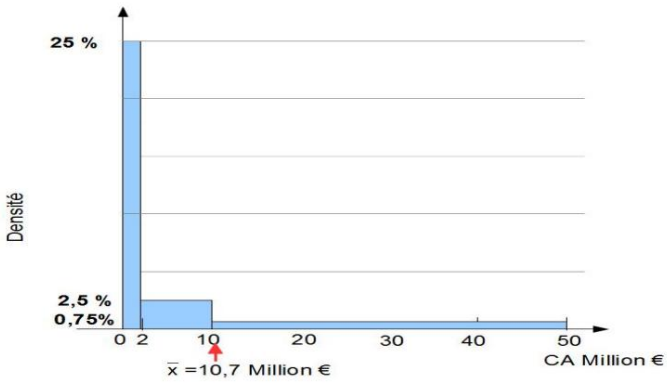
$\overline{x}$ = 0×0.33 +1×0.22 +2×0.28 + *3* ×0.11 +4×0.06 1.33

Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	4

Calculating the average with classes

If we only have the class sizes, we can calculate an approximate value of the mean by assigning to each class $[e_i, e_{i+1}[$ its **center** $c_i = \frac{e_i + e_{i+1}}{2}$.

These values are used instead of the terms in the formula.



Example. Among the SMEs in a region, we have

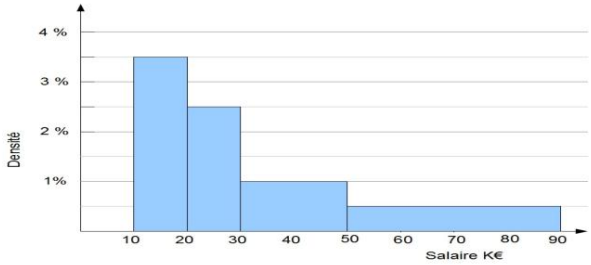
Annual turnover in millions of euros	Micro Enterprises [0,2[50% 0+2 [0,2[	Small Businesses [2,10[20% 2+10 [2,10[	Medium Enterprises [10,50[
Frequency	50%	20%	30%
Center	$\frac{0+2}{2} = 1$	$\frac{2+10}{2} = 6$	$\frac{10+50}{2} = 30$

$\bar{x} = 1 \times 0.5 + 6 \times 0.2 + 30 \times 0.3 = 10.7$

The average turnover of companies in the region is around 10.7 million euros.

Test

Here is the distribution by class of the annual salaries of the inhabitants of a region.



Annual salary Ke	[10,20[35%	[20,30[25%	[30,50[20%	[50,90[
Frequency				20%
Center	$\frac{10+20}{2} = 15$	$\frac{20+30}{2} = 25$	$\frac{30+50}{2} = 40$	$\frac{50+90}{2} = 70$

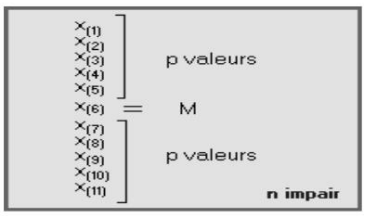
Mean $\bar{x} \simeq$

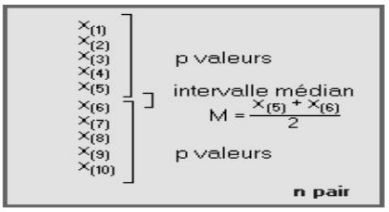
$$15 \times 0,35 + 25 \times 0,25 + 40 \times 0,2 + 70 \times 0,2 = 33,5 \text{ K€}$$

Median

The **median**

M is the middle value of the data series, that is, the value such that that there are as many observations “below” as “above”.



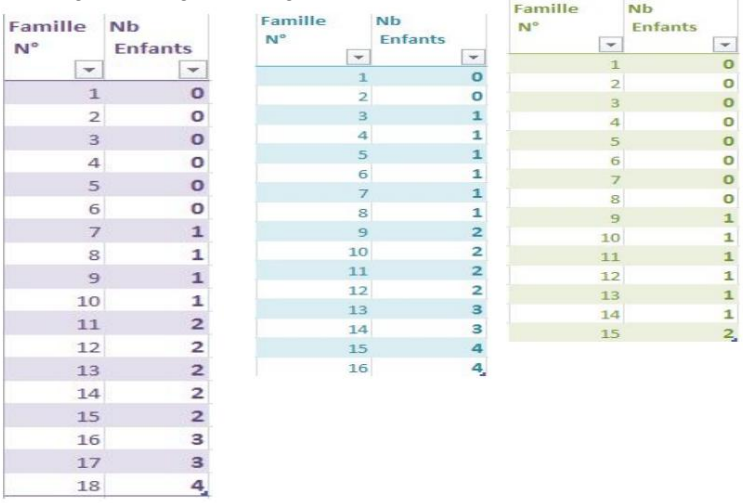


If the raw series of observed values is sorted in ascending order $x_1 \leq x_2 \leq \ldots \leq x_n$:

- if n is **odd,** let n = 2p + 1 then the median is $M = x_{p+1}$.
- if n is **even,** let n = 2p we generally choose $M = \frac{x_p + x_{p+1}}{2}$.

Example and Test

Building **A** Building **B** Building **C** Calculate the median number of children per family in the three buildings.



Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	4

Famille N°	Nb Enfants
1	0
2	0
3	1
4	1
5	1
6	1
7	1
8	1
9	2
10	2
11	2
12	2
13	3
14	3
15	4
16	4

Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	1
10	1
11	1
12	1
13	1
14	1
15	2

Building A :

MA = $\frac{x9+x10}{2}$ = $\frac{1+1}{2}$ = 1 because n = 18 = 2 × 9

Building B :

MB = $\frac{x8+x9}{2}$ = $\frac{1+2}{2}$ = *1.5*

because n = 16 = 2 × 8

Building C :

MC = x8 = 0

because n = 15 = 2 × 7 + 1

Median from cumulative numbers and frequencies

The median M can be obtained using the frequencies (or the numbers) cumulative:

the median is the first modality whose cumulative frequency is greater than 50% (or whose cumulative workforce is greater than half of total workforce).

Example. For the number of children in the 18 families studied,

Number of children xi	0	1	2	3	4
Cumulative workforce Ni	6	10	15	17	18
Cumulative frequencies Fi	33%	55%	83%	94%	100%

It is noted that 55% of families have at most one child and that 33% of families have no children, so the median is M = 1.

Similarly, “1” is the first value whose cumulative number (10) exceeds the half of half total effective $\frac{18}{2}$ = 9

Test

Here is the table of numbers and cumulative numbers concerning the variable "number of employees" on a population composed of 107 SMEs of a city.

Number of employees xi	0	1	2	3	4	5	6	Total
Effective nor	13	20	35	19	12	5	3	107
Cumulative workforce Ni	13	33	68	87	99	104	107	

The median is M= 2

Indeed = 53.5 $\frac{107}{2}$

The first cumulative workforce > 53.5 is 68, and this is the cumulative workforce of the modality 2.

Median from classes

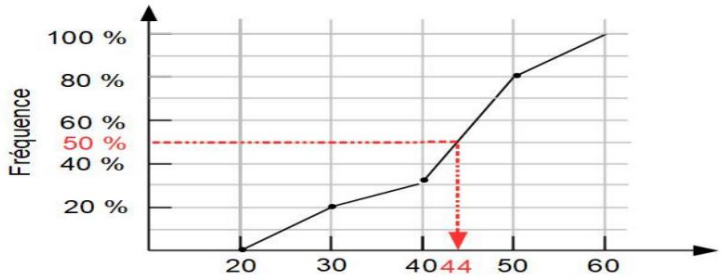
If we have the data by classes, we can obtain a value approximated the median using the cumulative frequency graph.

Example. The cumulative frequency table of the age of employees of a company is:

Age	[20,30[20%	[30.40[10%	[40,50[50%	[50,60[
Frequency				20%

Which gives the cumulative frequencies

Age	30	40	50	60
Cumulative frequencies 20% 30% 80% 100%				



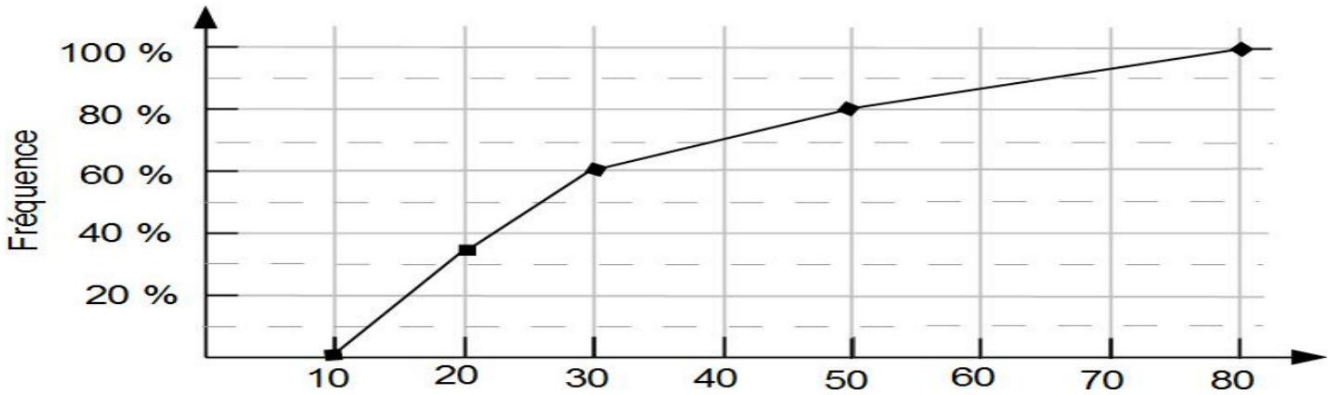
The graph shows that around 50% of employees have under 44 years old.

The median age is therefore M 44

Test

Here is the distribution by class of the annual salaries of the inhabitants of a region and its cumulative frequencies Annual salary Ke [10,20[

		[20,30[25%	[30.50[20%	[50,80[
Frequencies	35%			20%

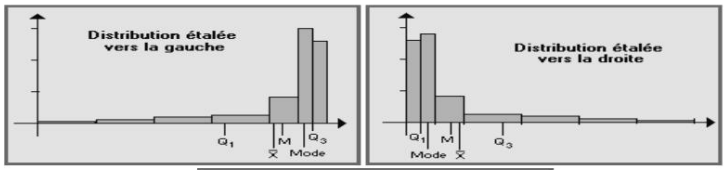


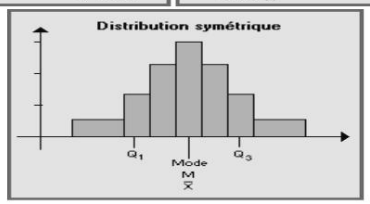
The median salary is M ȳ= 26 Ke

Median vs. Mean

In the case of a symmetric distribution, the median and mean coincide.

This is not generally true.





Distribution spread to the left:

$$\Rightarrow \overline{x} = \text{Mean} < M = \text{median}$$

Distribution spread to the right:

$$\Rightarrow M = \text{Median} < \overline{x} = \text{Mean}$$

Example: The distribution of wages is skewed to the right.

According to INSEE in France in 2013:

Median monthly salary = 1,772€ < 2,202€ = Average monthly salary.

Quartiles

Quartiles Q1, Q2 **and** Q3 are the three values that share the series 4-part statistic of the same workforce.

Q1 is the data in the series that separates the bottom 25% of the top 75%

Q2 is the data of the series which separates into two parts of same numbers. Q2 coincides with the median.

N°	Note	
1	2	
2	3	
3	4	
4	4	
5	6	Q1
6	6	
7	8	
8	9	
9	10	
10	10	Q2
11	10	
12	11	
13	11	
14	11	
15	11	Q3
16	12	
17	14	
18	14	
19	16	

Q3 is the data in the series that separates the bottom 75%
of the top 25%

Example. For the grades of the class of 19 students on the left,
Q1 = 6, Q2 = 10 and Q3 = 11.

The second quartile Q2 is the median.

Deciles and percentiles

Deciles D1, D2 .... and **D9** are the values that share the series
statistics in 10 games of the same workforce. Particularly significant are

D1 the data that separates the bottom 10% from the top 90%

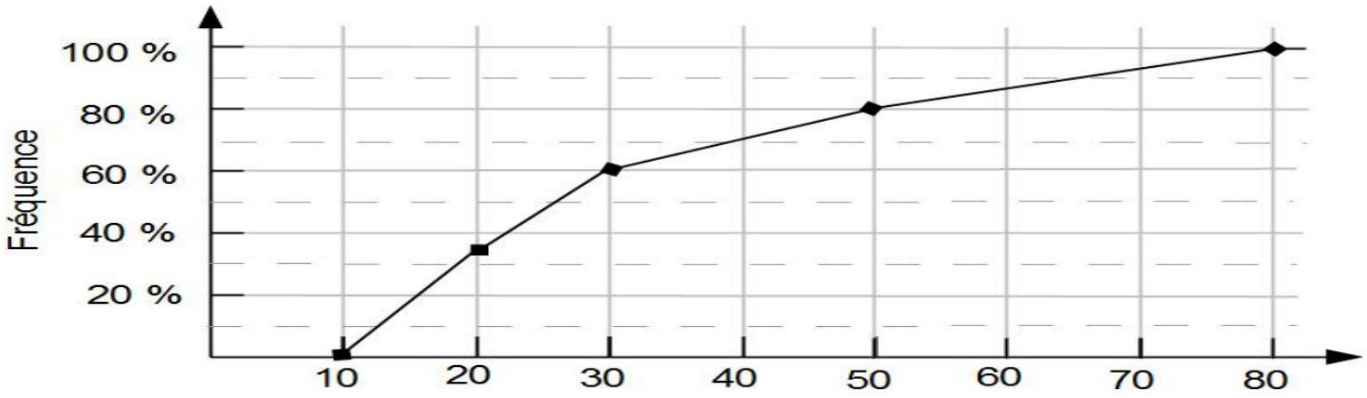
D9 the data that separates the bottom 90% from the top 10%

Example. According to INSEE for the monthly salary in France in 2013 we have
D1 = 1,200 and D9 = 3,544.
That is to say 10% of French people earn less than 1,200 euros and 10% more than
3,544 euros per month.

Test: from the cumulative frequency graph

Here is the distribution by class of annual salaries of the inhabitants of a
region and its cumulative

Annual salary K€	[10,20[	[20,30[	[30,50[	[50,80[
frequencies	35%	25%	20%	20%



Test: Fractiles of a discrete variable

Here is the table of workforce and cumulative workforce of the “Number of employees”
on a population composed of 107 SMEs in a city.

No. of employees xi	0	1	2	3	4	5	6	Total
Workforce ni	13	20	35	19	12	5	3	107
Cumulative workforce Ni	13	33	68	87	99	104	107	

Q1=1 and Q3= 3

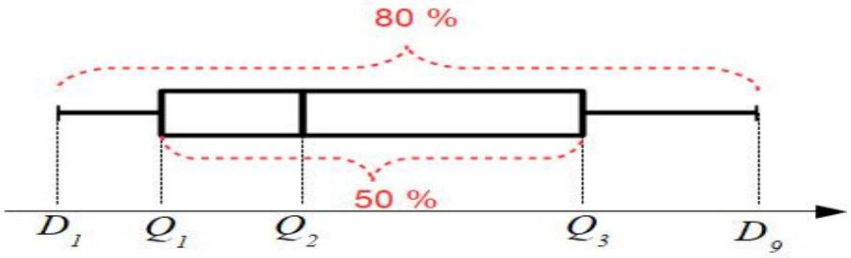
The first cumulative workforce > $107 \times \frac{1}{4}$ = 26.75 is the one associated with 1.
The first cumulative workforce > $107 \times \frac{3}{4}$ = 80, 25 is the one associated with 3.

D1= 0 and D9= 4

The first cumulative workforce > $107 \times \frac{1}{10}$ = 10, 7 is the one associated with 0.
The first cumulative workforce > $107 \times \frac{9}{10}$ = 96, 3 is the one associated with 4.

Box plot

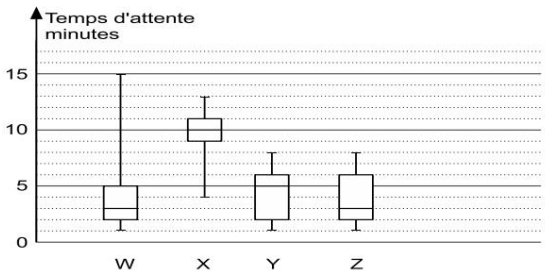
The **box plot** (or “box and whisker”) is used to represent the quartiles and deciles of a distribution.



The interest of these box diagrams is to be able to compare by a visual means several distributions of the same characteristic on different populations.

Test

A consumer association tested the telephone assistance services of four insurance companies W, X, Y and Z. For each call to the service, the customer's waiting time was measured.



- For the waiting time at **W customer service** we have:

  $D_1 =$ [1] $D_9 =$ [15]
  $Q_1 =$ [2] $Q_2 =$ [3] $Q_3 =$ [5]

- more than 15 min. and 50% more than 3 min.
  [ ] [ ]
- The worst customer service is X. [ ]
- The best is W or Z. [ ]