In this case the number of possible values is in principle infinite. It is rare that observations are repeated, because the greater or lesser precision of the measurements will only allow us to discern a finite number of distinct values.
A statistical series is always presented in ascending order.
To build the statistical table we must group the data in intervals called **classes**.

Description of classes

- These classes are of the form $[a_i, b_i[$ or $]b_i, a_i]$.
- Their number is chosen in an arbitrary way but very often close to

$$K = \begin{cases} \sqrt{n} \text{ if } n < 50 \\ 1 + \frac{10}{3} \log_{10} n \text{ if } n \geq 50 \end{cases}$$

- It is recommended to choose an odd number.
- These classes are disjointed and their union must contain all the data obtained.
- In this case we call n_i the number of observations belonging to the i^{th} class.
- The amplitude of a class (a_i, b_i) is the length of the interval $[a_i, b_i[$: $b_i - a_i$, such that

$$a = \frac{e}{K}.$$

Remark

It is often recommended to choose the same amplitude for all classes.

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

For our example we have $x_{\min} = 56$ et $x_{\max} = 90$ et $n = 30 < 50$, then

$$e = 90 - 56 = 34 \text{ and } K = \sqrt{30} = 5.48$$

we can take 5 or 6 classes. The amplitude of the classes is

$$a = \frac{e}{K} = \frac{34}{5} = 6.8 \approx 7 \text{ (or } a = \frac{34}{6} \approx 5.66 \approx 6)$$

then we can built the following classes
$[56, 62[$; $[62, 68[$; $[68, 74[$; $[74, 80[$; $[80, 86[$; $[86, 92[$ with $a = 6$, or
$[56, 63[$; $[63, 70[$; $[70, 77[$; $[77, 84[$; $[84, 91[$ with $a = 7$.

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers
$[56; 63[$	59.5	3
$[63; 70[$	66.5	2
$[70; 77[$	73.5	7
$[77; 84[$	80.5	14
$[84; 91[$	87.5	4
Total	–	30

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency
$[56; 63[$	59.5	3	0.1
$[63; 70[$	66.5	2	0.0667
$[70; 77[$	73.5	7	0.2333
$[77; 84[$	80.5	14	0.4667
$[84; 91[$	87.5	4	0.1333
Total	–	30	1

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency	%
$[56; 63[$	59.5	3	0.1	10
$[63; 70[$	66.5	2	0.0667	6.67
$[70; 77[$	73.5	7	0.2333	23.33
$[77; 84[$	80.5	14	0.4667	46.67
$[84; 91[$	87.5	4	0.1333	13.33
Total	–	30	1	100

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency	%	$n_i^c \nearrow$
$[56;63[$	59.5	3	0.1	10	3
$[63;70[$	66.5	2	0.0667	6.67	5
$[70;77[$	73.5	7	0.2333	23.33	12
$[77;84[$	80.5	14	0.4667	46.67	26
$[84;91[$	87.5	4	0.1333	13.33	30
Total	–	30	1	100	–

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency	%	$n_i^c \nearrow$	$n_i^c \searrow$
$[56;63[$	59.5	3	0.1	10	3	30
$[63;70[$	66.5	2	0.0667	6.67	5	27
$[70;77[$	73.5	7	0.2333	23.33	12	25
$[77;84[$	80.5	14	0.4667	46.67	26	18
$[84;91[$	87.5	4	0.1333	13.33	30	4
Total	–	30	1	100	–	–

Remark

x_i represents the center of the class $[a_i, b_i[$ (in physics we assimilate any object to its center of gravity). This allows us to return to the discrete case. We say that we discretize a continuous case.

Remark

x_i represents the center of the class $[a_i, b_i[$ (in physics we assimilate any object to its center of gravity). This allows us to return to the discrete case. We say that we discretize a continuous case.

Remark

The table does not correspond to the series obtained, but it has the advantage of better representing the entire population by eliminating the outliers.

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

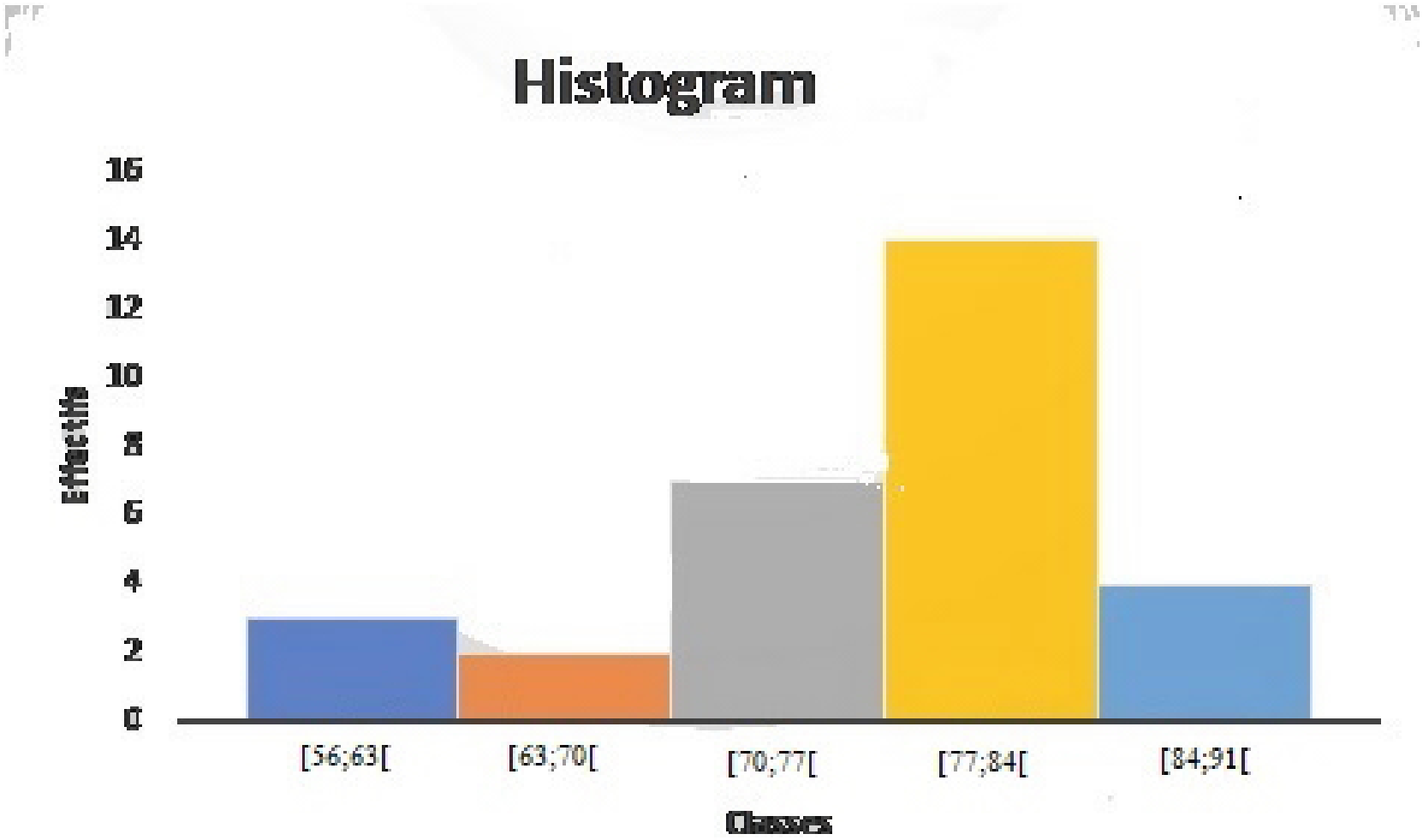
Histogram (case of classes of the same amplitude)

The histogram of a statistical series is the set of rectangles having for base the amplitude of a class and for height the number n_i of this class.

classes	Center x_i	Numbers	$n_i^c \nearrow$
$[56; 63[$	59.5	3	3
$[63; 70[$	66.5	2	5
$[70; 77[$	73.5	7	12
$[77; 84[$	80.5	14	26
$[84; 91[$	87.5	4	30
Total	–	30	–

Remark

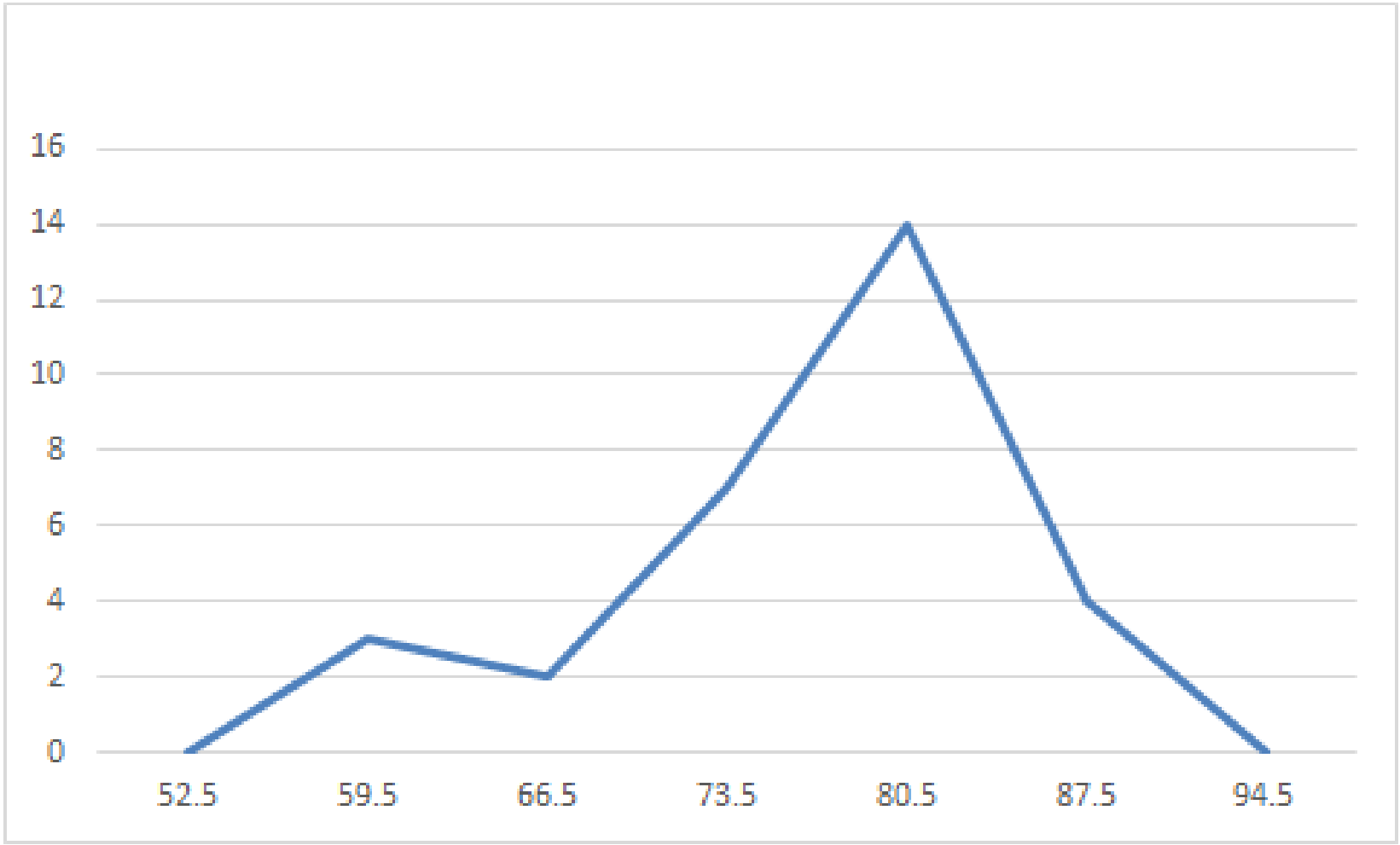
Histogram (case of classes with the same amplitude)



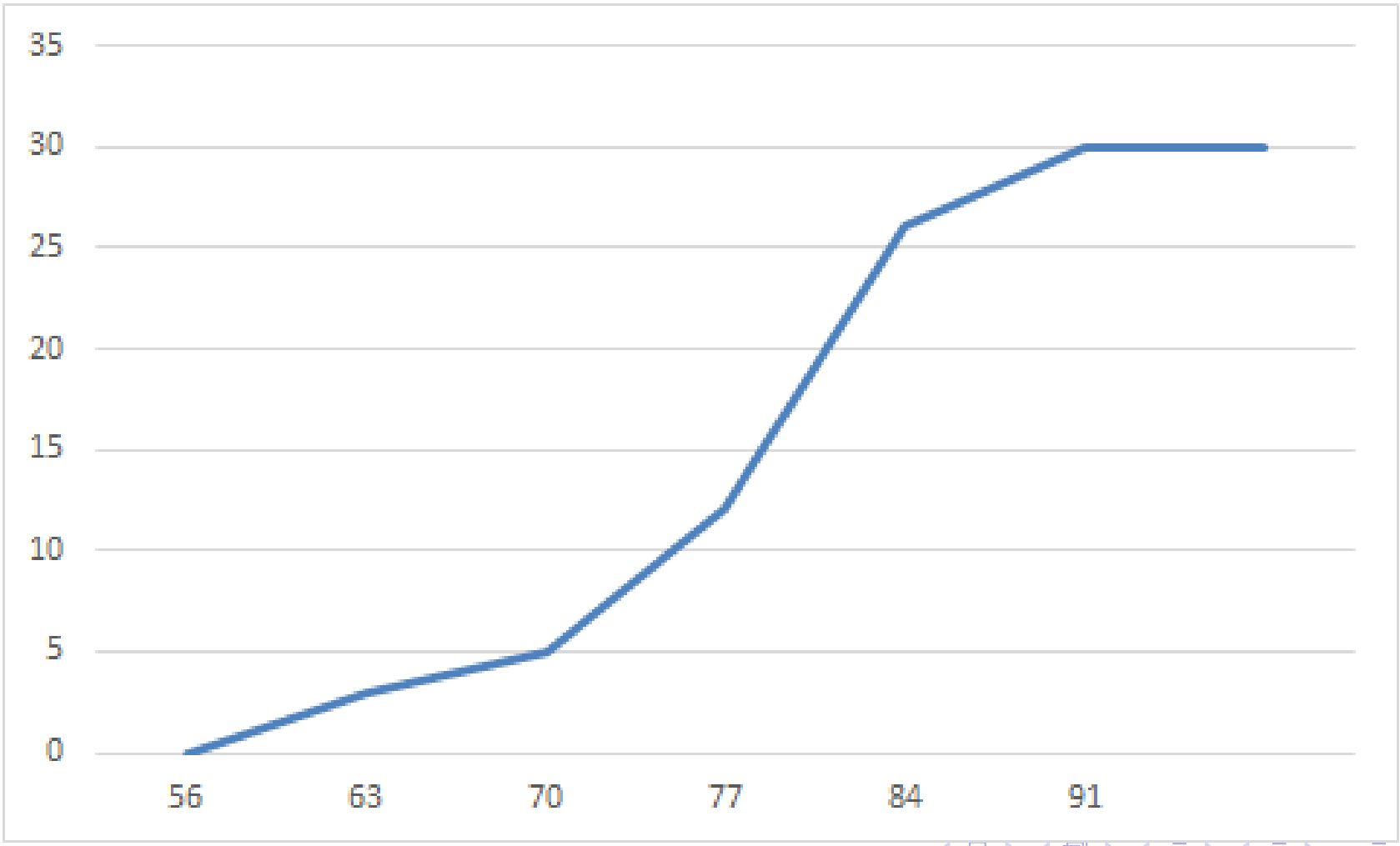
Histogram (case of classes of the same amplitude)
Remark: The total area of the Hitogram should always be equal to the sample size, considering the amplitude of the classes equal to the unit.
For the previous example we have

$$Area = (3 + 2 + 7 + 14 + 4) \times 1 = 30.$$

Absolute frequency polygon



Cumulative absolute frequency curve



Histogram in the case of classes of different amplitudes

Example (4)

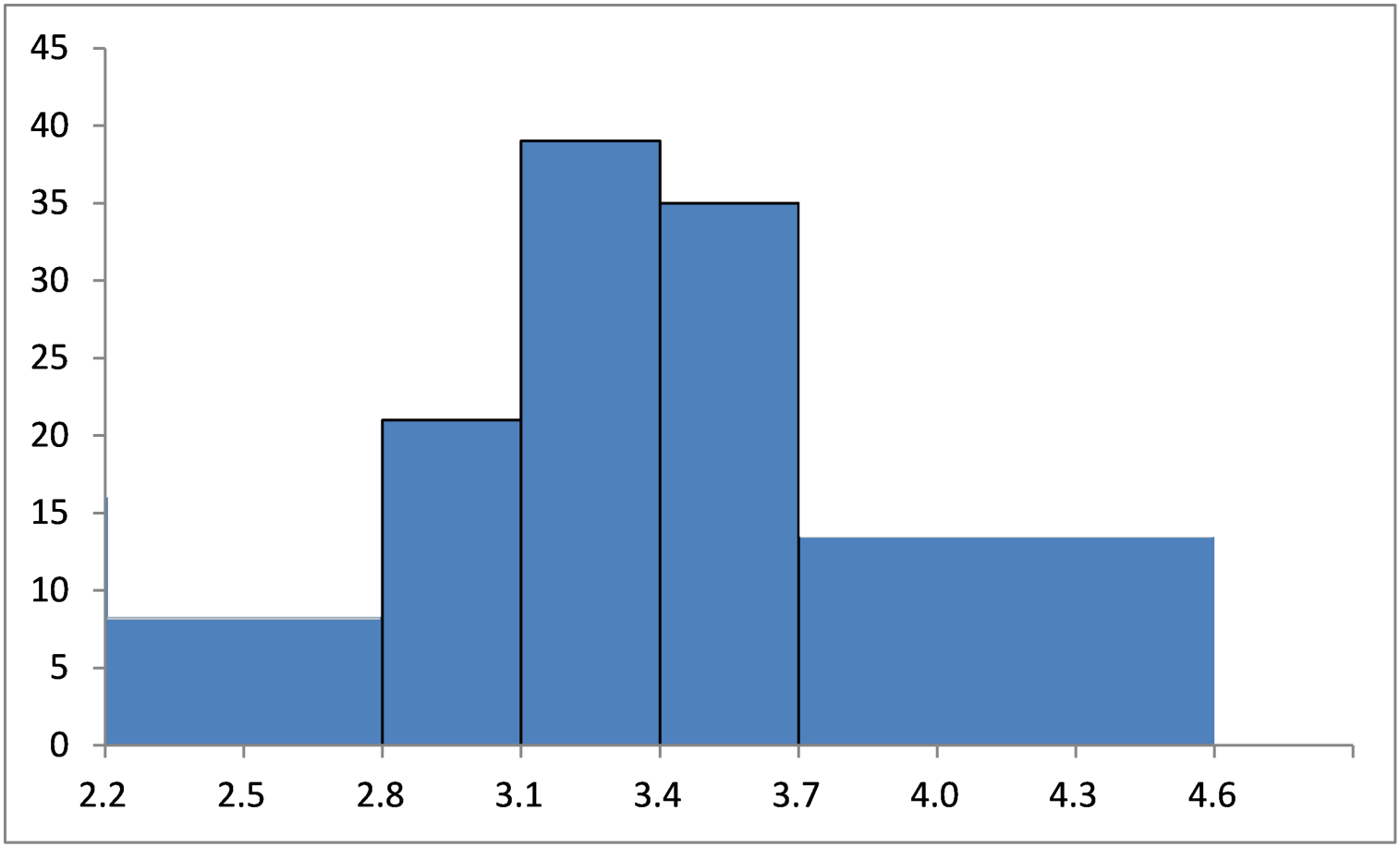
We consider the weight distribution in Kg of 150 newborns in a maternity hospital, the results are grouped in the following statistical table:

X Number of children	Center	Amplitude a	Numbers	Frequency
$[2.2; 2.8[$	2.50	$0,6$	16	0.1067
$[2.8; 3.1[$	2.95	$0,3$	21	0.1400
$[3.1; 3.4[$	3.25	$0,3$	39	0.2600
$[3.4; 3.7[$	3.55	$0,3$	35	0.2333
$[3.7; 4.6[$	4.15	$0,9$	39	0.2600
Total	–		150	1

Histogram in the case of classes with different amplitudes

To determine the histogram we determine the ratio between the different classes. For that, we consider the classes unit the classes having the smallest amplitude namely $[2.8; 3.1[$, $[3.1; 3.4[$ and $[3.4; 3.7[$ so we notice that the class $[2.2; 2.8[$ represent $\frac{0,6}{0,3} = 2$ unities and the classe $[3.7; 4.6[$ represent $\frac{0,9}{0,3} = 3$ unities. So to respect the previous remark we have to divide the number of students in the first class by 2 (i.e. $\frac{16}{2} = 8$) and the last class size by 3 (i.e. $\frac{39}{3} = 13$). We then obtain the following histogram:

Histogram in the case of classes with different amplitudes



Let's go back to example 3
first, we determine the modal class is $[77; 84[$ and then the mode will be the center of the class i.e. $Mo = 80,5$.
There is a second method which consists in estimating the mode by interpolation.
If $[a_i; b_i[$ is the modal class, then the mode is

$$Mo = a_i + \frac{e_1}{e_1 + e_2} a,$$

where
$a = b_i - a_i$ is the amplitude of the modal class;
e_1 is the difference between the numbers of the modal and previous classes;
e_2 is the difference between the numbers of the modal and following classes.

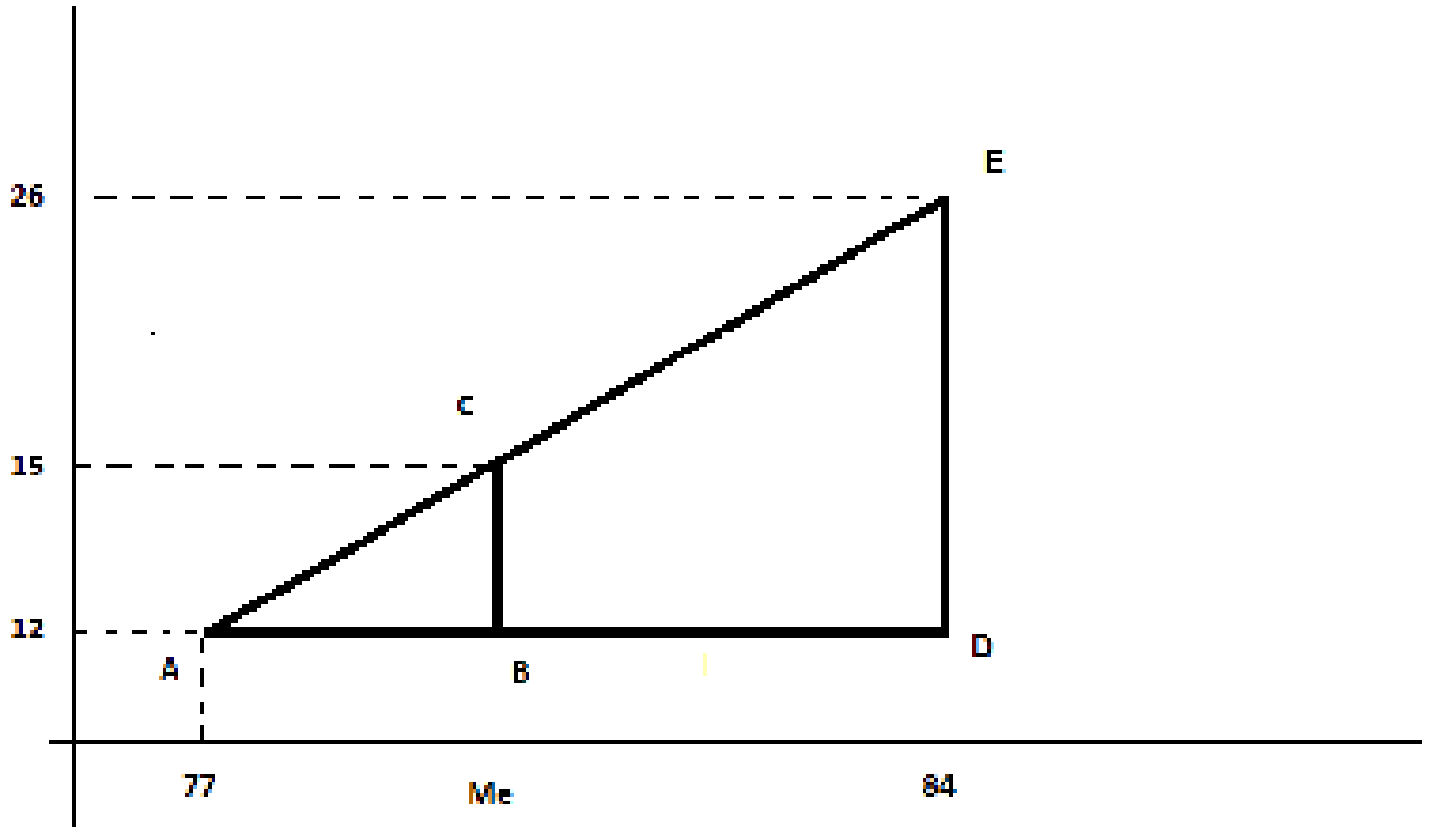
classes	Center x_i	Numbers	$n_i x_i$
$[56;63[$	59.5	3	178.5
$[63;70[$	66.5	2	133
$[70;77[$	73.5	7	514.5
$[77;84[$	80.5	14	1127
$[84;91[$	87.5	4	350
Total	–	30	2303

Then $\overline{X} = \frac{1}{30}\sum_{i=1}^{5} n_i x_i = \frac{2303}{30} \simeq 76.77.$

classes	Center x_i	Numbers	$n_i^c \nearrow$
$[56;63[$	59.5	3	3
$[63;70[$	66.5	2	5
$[70;77[$	73.5	7	12
$[77;84[$	80.5	14	26
$[84;91[$	87.5	4	30
Total	–	30	–

First, we determine the median class, which corresponds to half the numbers of people $\left(\frac{n}{2} = 15\right)$ hence $Me \in [77;84[$.

To determine *Me* we consider the curve of the increasing cumulative numbers but we will be interested only in the portion of the class $[77; 84[$



According to the theorem of Thales we have

$$\begin{aligned} \frac{AB}{AD} = \frac{BC}{DE} &\Longrightarrow \frac{Me-77}{84-77} = \frac{15-12}{26-12} \\ &\Longrightarrow Me = \frac{3}{14}7 + 77 \\ &\Longrightarrow Me = 78.5 \end{aligned}$$

Example

We take the example of the distribution of newborns according to their weight.

X : Number of children	Center	$n_i^c \nearrow$
$[2.2; 2.5[$	2.35	5
$[2.5; 2.8[$	2.65	16
$[2.8; 3.1[$	2.95	37
$[3.1; 3.4[$	3.25	76
$[3.4; 3.7[$	3.55	111
$[3.7; 4.0[$	3.85	131
$[4.0; 4.3[$	4.15	144
$[4.3; 4.6[$	4.45	150
Total	—	—

Determination of Q_1 : $\frac{n}{4} = 37.5 \leq 76$ then $Q_1 \in [3.1; 3.4]$ hence

$$Q_1 = \frac{37.5 - 37}{76 - 37} \times (3.4 - 3.1) + 3.1 = 3.104Kg$$

Determination of Q_3 : $\frac{3n}{4} = 112.5$ then $Q_3 \in [3.7; 4.0]$ hence

$$Q_3 = \frac{112.5 - 111}{131 - 111} \times (4.0 - 3.7) + 3.7 = 3.723Kg$$

hence

$$\begin{aligned} IQR &= Q_3 - Q_1 = 3.723 - 3.104 \\ &\Longrightarrow IQR = 0.619Kg. \end{aligned}$$

50% of the observations fall within an IQR length interval (between the values Q_1 and Q_3).

Example

We take the example of the distribution of newborns according to their weight.

X : Weight in Kg	x_i	n_i	$n_i^c \nearrow$	$n_i x_i$	$n_i x_i^2$
$[2.2; 2.5[$	$2,35$	5	5	$11,75$	$27,6125$
$[2.5; 2.8[$	$2,65$	11	16	$29,15$	$77,2475$
$[2.8; 3.1[$	$2,95$	21	37	$61,95$	$182,7525$
$[3.1; 3.4[$	$3,25$	39	76	$126,75$	$411,9375$
$[3.4; 3.7[$	$3,55$	35	111	$124,25$	$441,0875$
$[3.7; 4.0[$	$3,85$	20	131	77	$296,45$
$[4.0; 4.3[$	$4,15$	13	144	$53,95$	$223,8925$
$[4.3; 4.6[$	$4,45$	6	150	$26,7$	$118,815$
Total		150		$511,5$	$1779,795$

We have

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{8} n_i x_i = \frac{511,5}{150} \approx 3,41Kg \\
\sigma_X^2 &= \frac{1}{n}\sum_{i=1}^{8} n_i x_i^2 - \overline{X}^2 = \frac{1779,795}{150} - 3,41^2 \\
&\Longrightarrow \quad \sigma_X^2 \approx 0,2372 \Longrightarrow \sigma_X \approx 0,4870Kg.
\end{aligned}$$

And

$$\begin{aligned}
CV_X &= 100\frac{0,4870}{3,41} \\
&\Longrightarrow \quad 14,28\%.
\end{aligned}$$

We have

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{X}\right)^3}{\sigma_X^3}.$$

- The distribution is asymmetric if $|\gamma_1| > \frac{1}{2}$.
- The distribution spreads to the right if $\gamma_1 > 0$;
- The distribution spreads to the left if $\gamma_1 < 0$;
- The distribution is symmetric if $\gamma_1 = 0$;

We have

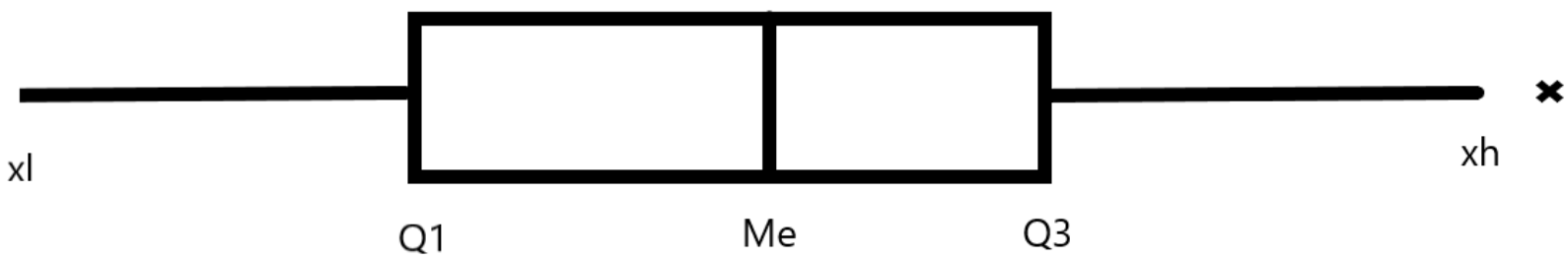
$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{X}\right)^4}{\sigma_X^4} - 3.$$

We say that

- the distribution is normal (the curve is mesokurtic) if $\gamma_2 = 0$;
- the distribution is less flattened than the normal distribution (the curve is leptokurtic) if $\gamma_2 > 0$;
- the distribution is more flattened than the normal distribution (the curve is platykurtic) if $\gamma_2 < 0$;

It is interesting to visualize concepts such as symmetry, dispersion or centrality of the distribution of values associated with a variable.
They are also very interesting to compare variables based on similar scales and to compare the values of observations of groups of individuals on the same variable.
To do that we want a graphical representation that answer all these questions and the box-plots is the method for graphically demonstrating the locality, dispersion and skewness groups of numerical data through their quartiles.
The box -plot or box-and-whisker plot was first introduced in 1970 by John Tukey, who later published on the subject in his book "Exploratory Data Analysis" in 1977.

1. We draw a rectangle of length IQR;
2. We locate the median by a line inside the box;
3. We calculate $Q_3 + 1,5 \cdot IQR$ and we look for the last observation $x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right)$;
4. We calculate $Q_1 - 1,5 \cdot IQR$ and look for the first observation $x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right)$;
5. We draw two lines from the midpoints of the rectangle widths to the values x_l and x_h which are called whiskers;
6. Any value that does not lie between the ends of the whiskers is usually represented by a star, which we will call the extreme value or outlier.


xl
Q1
Me
Q3
xh

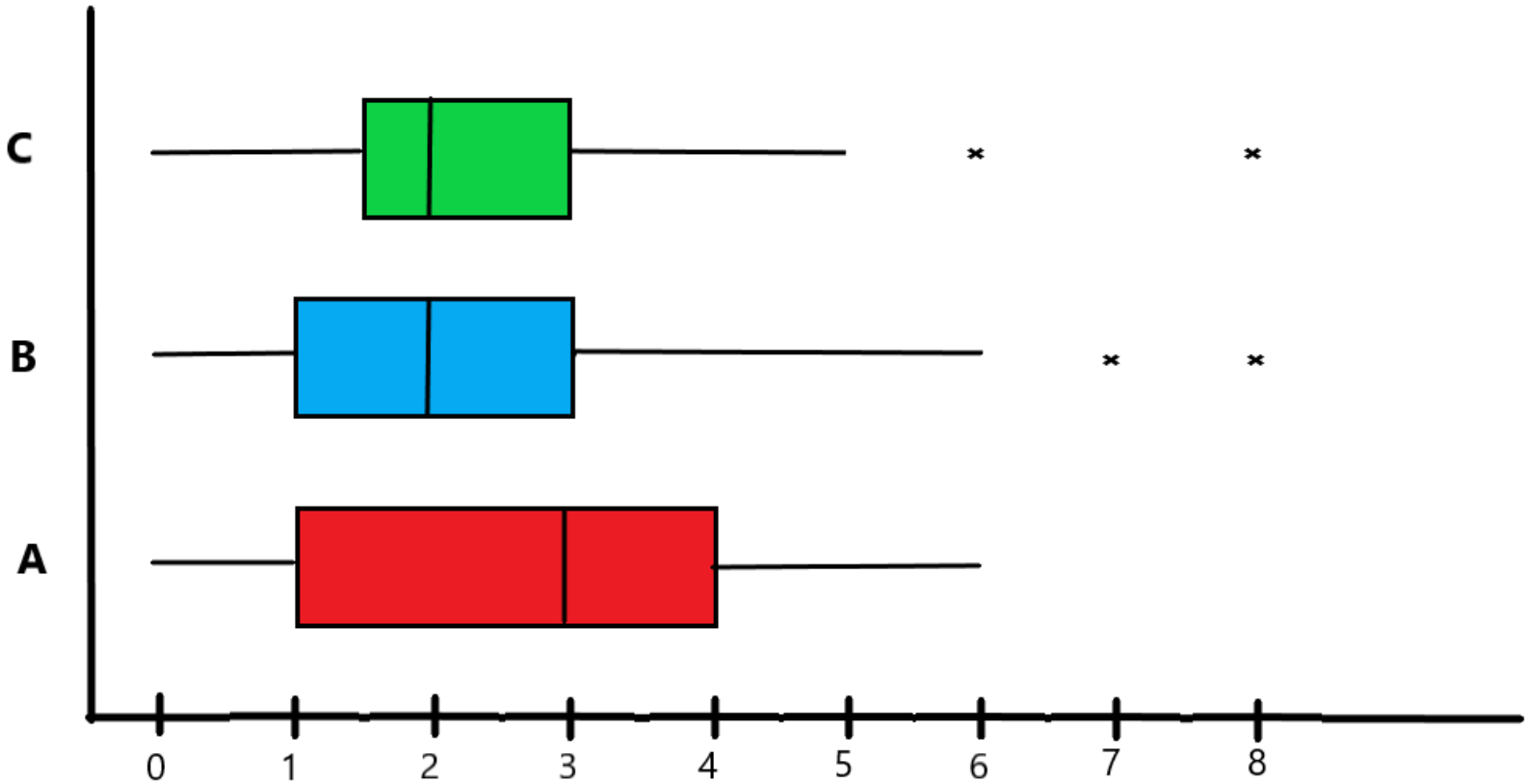
We consider the distribution of the number of children per family in three different neighborhoods A, B and C. Let X_A, X_B and X_C be the number of children per family in the neighborhoods A, B and C respectively.

X_A	n_i	$n_i^c \nearrow$
0	11	11
1	16	27
2	21	48
3	25	73
4	17	90
5	8	98
6	2	100
Total	100	

X_B	n_i	$n_i^c \nearrow$
0	12	12
1	28	40
2	46	86
3	38	124
4	20	144
5	2	146
6	1	147
7	1	148
8	2	150
Total	150	

X_C	n_i	$n_i^c \nearrow$
0	10	10
1	15	25
2	30	55
3	25	80
4	15	95
5	3	98
6	1	99
8	1	100
Total	100	

For the neighborhood A we have: $Me = 3; Q_1 = 1; Q_3 = 4$, then $IQR = 3$ so
$x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right) = \min\left(xi \middle| xi \geq -3,5\right) = 0$ and
$x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right) = \max\left(xi \middle| xi \leq 8,5\right) = 6.$
For the neighborhood B we have: $Me = 2; Q_1 = 1; Q_3 = 3$, then $IQR = 2$ so
$x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right) = \min\left(xi \middle| xi \geq -2\right) = 0$ and
$x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right) = \max\left(xi \middle| xi \leq 6\right) = 6$ and there is twi outliers 7 and 8.
For the neighborhood C we have: $Me = 2; Q_1 = 1,5; Q_3 = 3$, then $IQR = 1,5$ so
$x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right) = \min\left(xi \middle| xi \geq -1,25\right) = 0$ and
$x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right) = \max\left(xi \middle| xi \leq 5,25\right) = 5$ and there is two outliers 6 and 8.


C
B
A
0
1
2
3
4
5
6
7
8