PARIS PANTHÉON-ASSAS UNIVERSITÉ

Introduction to Data Analysis

Year 2025/2026 - Semester 4 - P2 - INT

Table des matières

Chapitre 1

Elements of descriptive statistics

1.1 Introduction

From the outset, human beings have stood out in the animal kingdom for their ability to collect and analyze data. As a result, we have gradually come to dominate our environment, becoming able to control it. Then, with the arrival of machines, the automation of data collection changed our relationship with the world, and data science came into being. In this course, we'll take a closer look at how they work.

In the Probability course, you saw the classic example of the 6-sided die. You know that there's a 1-in-6 chance of landing on a face, fixed in advance before the die is rolled.

In the Data Analysis course, we'll do things differently, not by working on the theoretical result of an infinite number of throws, but by rolling the die a finite number of times, for example, $n = 600$ times. Let's look at a summary of the results in the following table :

Face	1	2	3	4	5	6
Number of appearances	89	108	105	94	95	109

TABLE 1.1 – Results of 600 throws of a 6-sided die

While in probability we say that the random variable X follows a uniform distribution, here we see that the results observed per face differ slightly due to the finite number of throws. What's more, if we repeat the experiment, these values are likely to be different. What we can be sure of is that if n tends towards infinity, the distribution will tend towards a uniform distribution, in the case where the die is perfectly balanced. So several questions arise : how can we characterize the differences in each experiment ? And how do they evolve as a function of n ? We might also ask : how can we identify a slightly biased die from the results of 600 throws ? In this case, some faces seem to occur more than others. But what is the difference between the two, and can we say, with a small risk of being wrong, that the die is biased or not ? This question brings us back to the question of a threshold to be exceeded in order to

reach a conclusion. Naturally, we'd like this threshold to be the same for every scientist on the planet, as science is intended to be universal.

What we've just seen is the core of the **scientific method** for accessing knowledge, and has been discussed by philosophers since Antiquity. In this course, we will first learn how to describe data, then how to analyze a large amount of data, and finally how to run tests on data.

1.2 Data

Définition 1.1 (Individuals and population).
Let be a set of **individuals** whose union is said to form a **population**. The total number of individuals in this population is called **total number**. When the study of a population is not possible because the number of individuals is too large, we work on a **sample** of n individuals.

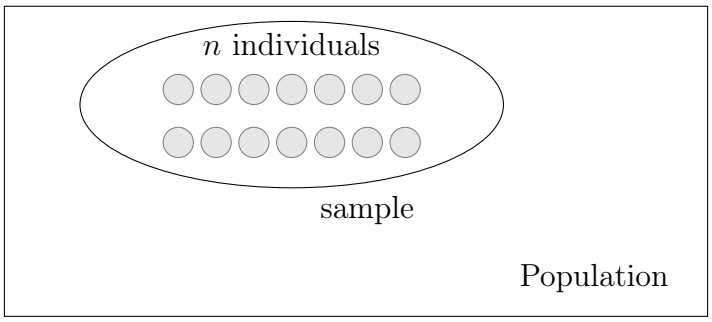


FIGURE 1.1 – The n individuals constitute a sample in a population.

For a statistical study to be of high quality, the sample, i.e. the part of the population tested, must be **representative** of the population. For a sample to have the best possible chance of representing our population, we need to draw n individuals from the sample **randomly**. We'll see later that, even if the sample is chosen at random, there's always a risk that it won't represent our population well.

To do this, we can imagine a poll of 1,000 people for a presidential election. Let's draw a sample of 1,000 people at random from the entire French population. There is always a risk that, through bad luck, we will systematically find people voting for the same candidate in the election. So it's not always easy to draw up and work on samples. We'll come back to this when

we discuss statistical tests.

Définition 1.2 (A realization of a sample).
Consider n random variables, X_k for $k \in [\![1, n]\!]$, independent and of the same law X. A **realisation of a sample** of size n is a tuple $(x_1, x_2, \ldots, x_n)$ of values taken by the sample : this is our collected data.

Définition 1.3 (Statistical data).
We take two types of data from the individuals in our sample :

- Qualitative data are also known as categorical data : they are discrete and concern categories to which we can associate a name. The various possible names are the **modalities** of the variable.
- Data **quantitative** are numerical values : they can be discrete or continuous.

Examples of qualitative data : the eye color, type of transport taken this morning, brand of shoes, etc.

Sometimes in qualitative data, an **order** can be assigned. For example, the ratings "Fair", "Good", "Very good", "Excellent" attributed to a DM have an obvious order. This is referred to as a **ordinal** qualitative variable.

Examples of quantitative data : height, DE score from "Functions of several variables", etc.

1.3 One-variable statistics

In the first instance, we take only one data item per individual, so our statistics are univariate.

1.3.1 Central tendency parameters

Définition 1.4 (Empirical average).
Consider a population of individuals and a sample of n individuals. Statistical data are taken according to a random variable X for each individual, noted x_i for individual number $i \in [\![1, n]\!]$. The quantity :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is called **empirical average**, denoted $\overline{x}$.

Exemple: Given a class of $n = 3$ students, we consider the DE grade given for the first semester's MF. If these students' marks are $x_1 = 8$, $x_2 = 10$ and $x_3 = 14$, then the average for this class is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1+x_2+x_3}{n} = \frac{32}{3} \simeq 10.67$

Be careful with significant figures, if you're asked for them!

Définition 1.5 (Median).
The value separating the lower half from the upper half of the terms of an ordered quantitative statistical series is called the **median**, denoted Me. It can also be defined for an ordinal qualitative variable.

Exemple: A class of 40 students graduate from Efrei, and their salaries are noted. 39 students leave the school with a monthly salary of 2500 euros, and one student with a monthly salary of 1 million euros. Calculating the average gives $\overline{x} \simeq 27438$ €. We agree that this value in no way reflects what's going on in the classroom. On the other hand, half the students have less than 2500, so the median is 2500, which is a relevant value for describing the center, i.e. the main information, of a distribution. Compared with the average, the median is insensitive to extreme values, but its calculation is a little more complex.

Point méthode 1.1: Median calculation

- Classify values.
- If n is even, the median is the average between the values of the ranks $\frac{n}{2}$ and $\frac{n}{2}+1$.
- If n is odd, the median is the value at rank $\frac{n+1}{2}$.

Exemple:

Let's consider the 5 values : $x_1 = 12$, $x_2 = 10$, $x_3 = 12$, $x_4 = 20$, $x_5 = 7$.
We first classify the values : 7, 10, 12, 12, 20.
$n = 5$ is odd, so the median is $(5 + 1)/2 = 3$. It's the third value, i.e. 12, that we see in the middle of these ranked values.

Exemple:

Now consider the 6 values : $x_1 = 12$, $x_2 = 10$, $x_3 = 12$, $x_4 = 20$, $x_5 = 7$, $x_6 = 4$.
We classify the values : 4, 7, 10, 12, 12, 20

$n = 6$ is even, so the median is the average of the values at ranks $6/2 = 3$ and $6/2 + 1 = 4$, i.e. the average of the third and fourth values, i.e. the average of 10 and 12, i.e. 11.

Définition 1.6 (Quartile).
The **first quartile** Q_1 is the value separating the lower quarter of the terms of an ordered quantitative statistical series.
The **third quartile** Q_3 is the value separating the lower three-quarters of the terms in an ordered quantitative statistical series.

Point méthode 1.2: Calculating quartiles

- Rank the values
- If n is a multiple of 4, Q_1 is the average between the values ranked $\frac{n}{4}$ and $\frac{n}{4} + 1$. And Q_3 is the average between values ranked $\frac{3n}{4}$ and $\frac{3n}{4} + 1$.
- If n is not a multiple of 4, Q_1 is the value at rank $E(\frac{n}{4}) + 1$ and Q_3 is the value at rank $E(\frac{3n}{4}) + 1$, with E the integer function.

Exemple: Let's consider 8 values : $x_1 = 12$, $x_2 = 10$, $x_3 = 12$, $x_4 = 20$, $x_5 = 7$, $x_6 = 4$, $x_7 = 1$, $x_8 = 5$.

We classify the values : 1 ; 4 ; 5 ; 7 ; 10 ; 12 ; 12 ; 20.

Since n is a multiple of 4, the first quartile is the average of the values in the ranks $\frac{n}{4} = 2$ and $\frac{n}{4} + 1 = 3$, i.e. the average of the second and third values, i.e. the average of the values 4 and 5, i.e. 4.5. Similarly, the third quartile is 12.

Exemple: Let's consider 9 values : x_1 =12, x_2 =10, x_3 =12, x_4 =20, x_5 =7, x_6 =4, x_7 =1, x_8 =5, x_9 =20.

We classify the values : 1 ; 4 ; 5 ; 7 ; 10 ; 12 ; 12 ; 20 ; 20.

Since n is not a multiple of 4, the first quartile is the value at rank $E(\frac{n}{4}) + 1 = E(2,25) + 1 = 2 + 1 = 3$, i.e. the third value, 5. Similarly, the third quartile is at rank 7, i.e. the value 12.

Définition 1.7 (Mode).
The most represented value in a sample is called **mode**, or dominant value, noted Mo.

Définition 1.8 (Frequency).
The **frequency** is the value $f_i = \frac{n_i}{n}$ where n_i is the number of times the value is present in the sample of size n, which is the total number.

Exemple:

Given 10 values :

$x_1 = 12$, $x_2 = 12$, $x_3 = 12$, $x_4 = 20$, $x_5 = 20$, $x_6 = 4$, $x_7 = 4$, $x_8 = 4$, $x_9 = 4$, $x_{10} = 4$.

The value 4 appears the most times, so 4 is the mode. Its frequency is $f = \frac{5}{10} = 0.5 = 50\%$, as it appears 5 times out of the ten values.

The frequency of the value 12 is $f = \frac{3}{10} = 0.3 = 30\%$.

The frequency of the value 20 is $f = \frac{2}{10} = 0.2 = 20\%$.

We can draw the **bar graph** and the **circular graph** :

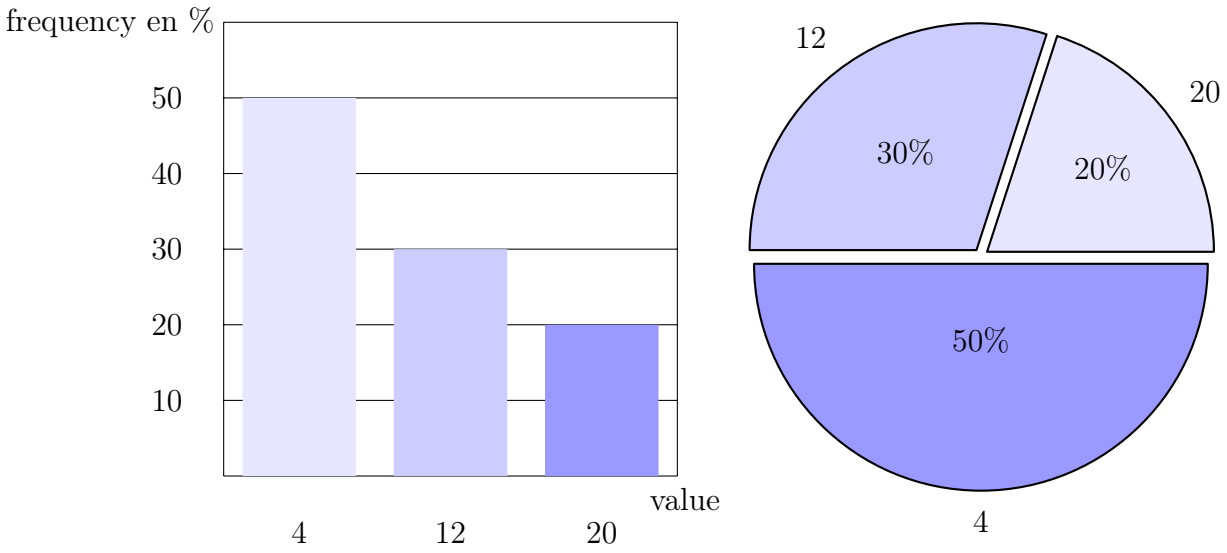


FIGURE 1.2 – Bar and pie charts

1.3.2 Dispersion parameters

Définition 1.9 (Observed or empirical standard deviation).
The quantity is called the observed or empirical variance of the series of sample values :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \overline{x}^2$$

The quantity σ is called **observed standard deviation** of the variable X in the sample. The x_i values, the average $\bar{x}$ and the standard deviation σ have the same units.

REMARK : : If, instead of considering a grade as a random variable, you take an individual's money in his or her Livret A passbook, then the average will have x_i as its unit, as will x_i and the standard deviation σ.

The standard deviation measures the **dispersion** of a variable, i.e. the extent to which x_i are spread out. The "smaller" σ is, the more "concentrated" the X values are around $\overline{x}$. In the extreme case of an entire sample having the same value, the standard deviation would be zero.

Démonstration.

$$\begin{aligned}\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 =& \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \overline{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\sum_{i=1}^{n}(2\bar{x}x_i) + \frac{1}{n}\sum_{i=1}^{n}(\overline{x}^2) \\ &= \overline{x^2} - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}(x_i) + \frac{1}{n}n\overline{x}^2 \\ &= \overline{x^2} - 2\bar{x}\bar{x} + \overline{x}^2 \\ &= \overline{x^2} - \overline{x}^2\end{aligned}$$

□

Exemple:

Let's take the previous example of our class of $n = 3$ students, each with a score on the first-semester MF, the students' scores being $x_1 = 8$, $x_2 = 10$ and $x_3 = 14$.

$\sigma^2 = \overline{x^2} - \overline{x}^2 = \frac{1}{3}(8^2 + 10^2 + 14^2) - \left(\frac{32}{3}\right)^2 = \frac{56}{9}$

Proposition 1.3

Let be a sample realization of the random variable X and let be $(\alpha, \beta) \in \mathbb{R}^2$:

$$\begin{aligned}\overline{\alpha x + \beta} =& \alpha\overline{x} + \beta \\ \sigma_{\alpha x+\beta} =& \alpha^2\sigma_x\end{aligned}$$

Démonstration.

$$\overline{\alpha x + \beta} = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta) = \frac{\alpha}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\beta = \alpha\overline{x} + \frac{1}{n}n\beta = \alpha\overline{x} + \beta$$

$$\sigma^2_{\alpha x+\beta} = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta - (\overline{\alpha x + \beta}))^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta - \alpha\overline{x} - \beta)^2 = \frac{\alpha^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \alpha^2\sigma_x^2$$

□

Exemple: Let's take the previous example of our class of $n = 3$ students, each with a score on the first-semester MF, the students' scores being $x_1 = 8$, $x_2 = 10$ and $x_3 = 14$. You can practice

by calculating the values of $3X + 1$ in the sample, then calculating the average, which would be 33, and the variance, which would be 56.

Définition 1.10 (Range).
The **range** of a sample of size n drawn from a random variable X is defined as the difference between the largest and the smallest values in the sample.

Définition 1.11 (Centered reduced statistical variable).
We call **centered reduced random variable**, the random variable Z calculated thanks to any random variable X, such that :

$$Z = \frac{X - \overline{x}}{\sigma_x}$$

with σ_x the observed or empirical standard deviation and $\overline{x}$ the average of the random variable X in the sample. The random variable Z then has a zero average and a standard deviation 1.

Démonstration.

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sigma_x} = \frac{1}{n\sigma_x}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{n\sigma_x}\left(\left(\sum_{i=1}^{n} x_i\right) - n\overline{x}\right) = \frac{1}{n\sigma_x}(n\overline{x} - n\overline{x}) = 0$$

$$\sigma_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_x}\right)^2 = \frac{1}{\sigma_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{\sigma_x^2}\sigma_x^2 = 1$$

□

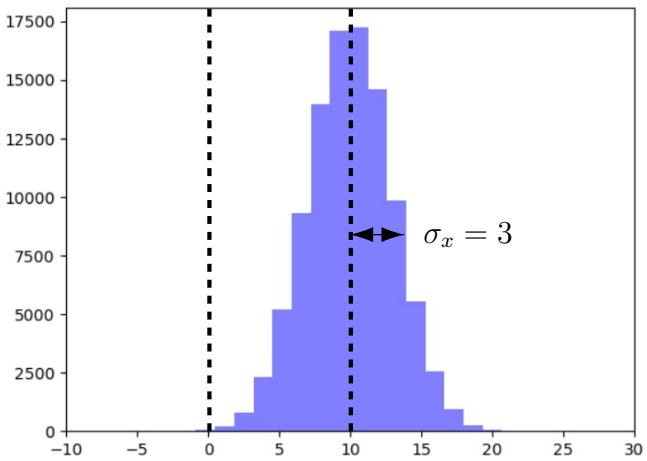
Exemple: Using Python, we want to generate 10,000 draws according to a Gaussian random variable X with $\overline{x} = 10$ and $\sigma_x = 3$. Then we'd like to count the numbers in 20 intervals (classes) between the largest and smallest values. And form a histogram.

To do this, we proceed in reverse, generating a random draw according to a centered reduced normal distribution, then changing the variance and finally changing the average. Let's take a look at the code :

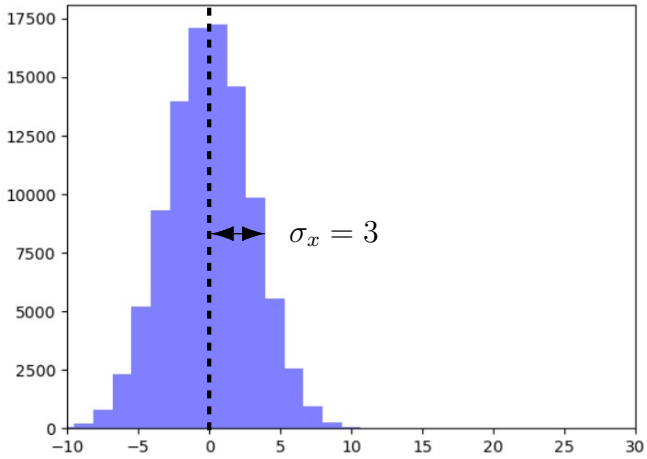
Code Python 1.3

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

```

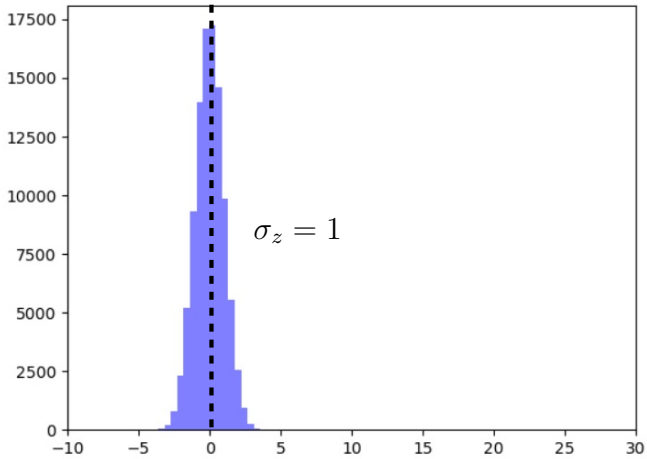
```
#Generation of the random draw according to a centered reduced normal
    distribution
N = 100000
x = stats.norm.rvs(size=N)
num_bins = 20
plt.hist(x, bins=num_bins, facecolor='blue', alpha=0.5)

y = np.linspace(-4, 4, 1000)
bin_width = (x.max() - x.min()) / num_bins

plt.xlim(-10, 30)
plt.show()

#Same random draw but with a different variance
plt.hist(3*x, bins=20, facecolor="blue", alpha=0.5)

plt.xlim(-10, 30)
plt.show()

#Same random draw but with different variance and average
plt.hist(3*x + 10, bins=20, facecolor="blue", alpha=0.5)

plt.xlim(-10, 30)
plt.show()
```



Variable X



Variable $X - \overline{x}$



Variable $Z = \frac{X-\overline{x}}{\sigma_x}$

FIGURE 1.3 – Histograms plotted with Python for 10000 draws according to a Gaussian X random variable of $\overline{x} = 10$ and $\sigma_x = 3$, then centered and reduced. The histogram is composed of 20 intervals between the largest and smallest values.

REMARK :
- The advantage of centering and reducing variables is to work with unitless variables, varying with comparable amplitudes around 0.
- On a Gaussian distribution, the standard deviation is the distance between the average and the inflection point.

Définition 1.12 (Asymmetry and kurtosis coefficients).
The asymmetry (skewness) coefficient (S_k) and the kurtosis coefficient (K) for a realization of sample size n on a random variable X are :

$$S_k = \frac{1}{\sigma_x^3} \sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^3 \qquad \text{and} \qquad K = \frac{1}{\sigma_x^4} \sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^4$$

with σ_x the observed or empirical standard deviation and $\bar{x}$ the average of the random variable X in the sample.

Proposition 1.4

The coefficient S_k evaluates the lack of symmetry of a distribution. It is zero for a symmetrical distribution (e.g. a normal distribution, or a binomial distribution with p=0.5). It is positive for a "right-spread" distribution.

Exemple: We generate 10000 draws according to a random variable $\mathcal{B}(n = 12, p = 0,7)$:

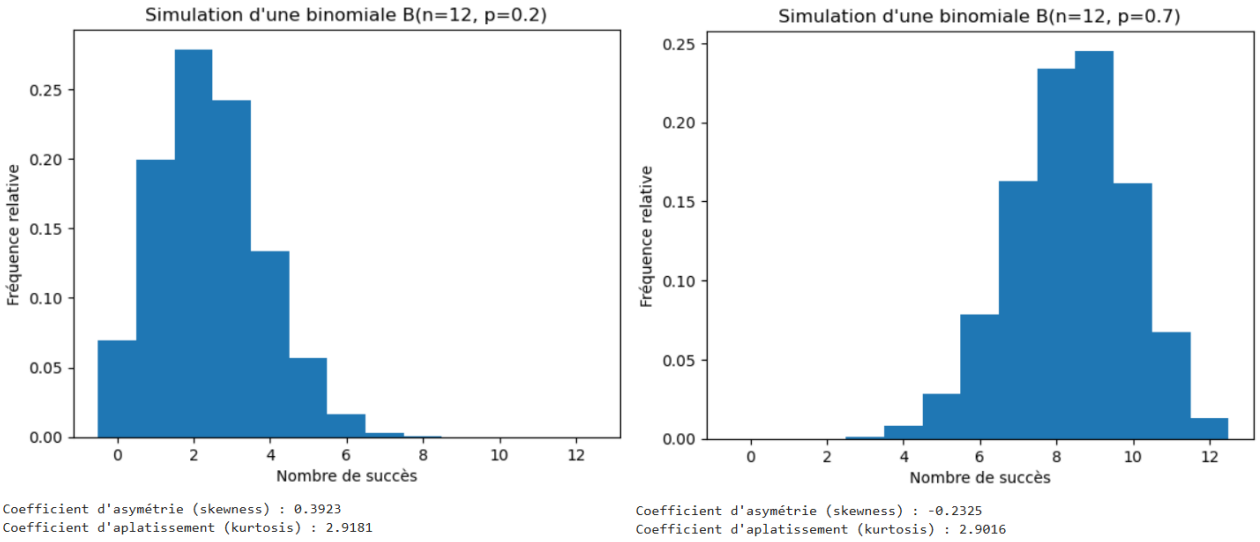
Code Python 1.4

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom, skew, kurtosis

n = 12
p = 0.7
size = 10000  # Sample size

samples = np.random.binomial(n, p, size)
plt.hist(samples, bins=np.arange(-0.5, n+1.5, 1), density=True)
plt.title(f"Binomial simulation B(n={n}, p={p})")
plt.xlabel("Number of successes")
plt.ylabel("Relative frequency")
plt.show()
```

And to calculate the statistical coefficients, we add :

Code Python 1.4

```
# Statistical coefficients from simulated samples
Asymmetry = skew(samples)
flattening = kurtosis(samples, fisher=False)
print(f"asymmetry coefficient (skewness) : {asymmetry:.4f}")
print(f"kurtosis coefficient : {flattening:.4f}")
```



Histogram on the right : distribution of a $\mathcal{B}(n = 12, p = 0.7)$ distribution, we obtain $S_k = -0.23$ and $K = 2.9$. It is well spread out on the left.

Histogram on the right : distribution of a $\mathcal{B}(n = 12, p = 0.7)$ distribution, we obtain $S_k = -0.23$ and $K = 2.9$. It is well spread out on the left.

FIGURE 1.4 – Histogram on the left : distribution of a $\mathcal{B}(n = 12, p = 0,2)$ law, we obtain $S_k = 0,39$ and $K = 2,9$. It is well spread out on the right.
Histogram on the right : distribution of a $\mathcal{B}(n = 12, p = 0.7)$ distribution, we obtain $S_k = -0.23$ and $K = 2.9$. It is well spread out on the left.

1.4 Two-variable statistics

Let n be the individuals in a sample from which, this time, two values are taken for each. We have a bivariate statistical series.

1.4.1 Definitions

Définition 1.13 (Point clouds).
Let a sample of size n be drawn from two random variables X and Y. The **point cloud** is the representation of the n points with abscissa x_i and ordinates y_i, for $i \in \{1, \dots, n\}$.

Définition 1.14 (Centroid).
Let a sample of size n be drawn from two random variables X and Y. The **centroid** of a point cloud is the point with coordinates $(\overline{x}, \overline{y})$.

Définition 1.15 (Covariance).
Consider a realization of sample size n on two random variables X and Y. We call the **covariance** the value :

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x - \overline{x})(y - \overline{y}) = \overline{xy} - \overline{x}, \overline{y}$$

.

This is an extension of the notion of variance. It quantifies their joint deviations from their respective averages. Note that $\sigma_{xx} = \sigma_x^2$.

Démonstration.

$$\begin{aligned}\sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x - \overline{x})(y - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(xy - \overline{x}y - x\overline{y} + \overline{x}\,\overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} xy - \overline{x}\frac{1}{n}\sum_{i=1}^{n} y - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x + \frac{1}{n}\sum_{i=1}^{n}\overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y} - \overline{y}\,\overline{x} + \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y}\end{aligned}$$

□

Définition 1.16 (Correlation coefficient).
Consider a sample realization of size n on two random variables X and Y such that their variances are non-zero (i.e. the variables are not constants). We call the **coefficient of correlation between X and Y** the value :

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Proposition 1.5

The correlation coefficient rho_{xy}, calculated on a realization of the sample on two random variables X and Y satisfies :

$$-1 \leq \rho_{xy} \leq 1$$

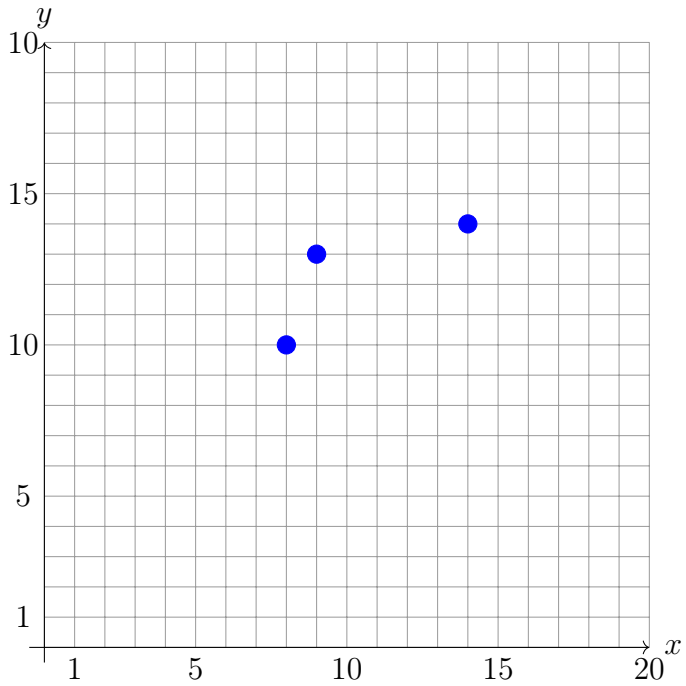
Démonstration. Using the Cauchy-Schwarz inequality, we verify that :

$$\begin{aligned} |\sigma_{xy}| &= \left| \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y - \overline{y}) \right| \\ &= \frac{1}{n} < x - \overline{x}, y - \overline{y} > \\ &\leq \frac{1}{n} \|x - \overline{x}\| . \|y - \overline{y}\| = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \sigma_x \sigma_y \end{aligned}$$

Then $-1 \leq \frac{\sigma_{xy}}{\sigma_x \sigma_y} \leq 1$

Furthermore, the Cauchy-Schwarz inequality becomes an equality if and only if $x_i - \overline{x} = a(y_i - \overline{y})$, which would imply that the two statistical series would be linked by an affine relationship. □

Exemple: Let's consider three students. Each had a grade in the DE "Multi-variable Functions" and a grade in the DE "Probability" in the first semester : X is the random variable for the grade in the "Multi-variable Functions" course and Y is the random variable for the grade in the "Probability" course. The first student had $(x_1, y_1) = (8, 10)$, the second $(x_2, y_2) = (9, 13)$ and the third $(x_3, y_3) = (14, 14)$. Let's plot the point cloud :



We compute the average of the students on the variable X :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3}{n} = \frac{31}{3}.$$

And on the variable Y :

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + y_3}{n} = \frac{37}{3}.$$

The centroid of the point cloud is located at the point $(\overline{x}, \overline{y}) = \left(\frac{31}{3}, \frac{37}{3}\right)$. $(\overline{x}, \overline{y}) = (\frac{31}{3}, \frac{37}{3})$

We can compute the sample variances of the two variables in the dataset : $\sigma_x^2 = \overline{x^2} - \overline{x}^2 = \frac{62}{9}$ and $\sigma_y^2 = \overline{y^2} - \overline{y}^2 = \frac{26}{9}$. The covariance is : $\sigma_{xy} = \overline{xy} - \overline{x}\,\overline{y} = \frac{32}{9}$.

The correlation coefficient is : $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{32}{\sqrt{62}\sqrt{26}} \simeq 0,8$

In the following figure, we can observe how the correlation coefficient behaves for different types of point clouds. Correlation is very often reduced to linear correlation between quantitative variables, that is, the adjustment of one variable with respect to the other through an affine relationship. While this interpretation can be kept in mind, it is also important to note that symmetric distributions can produce a zero correlation coefficient even when a clear relationship between the variables is visible.

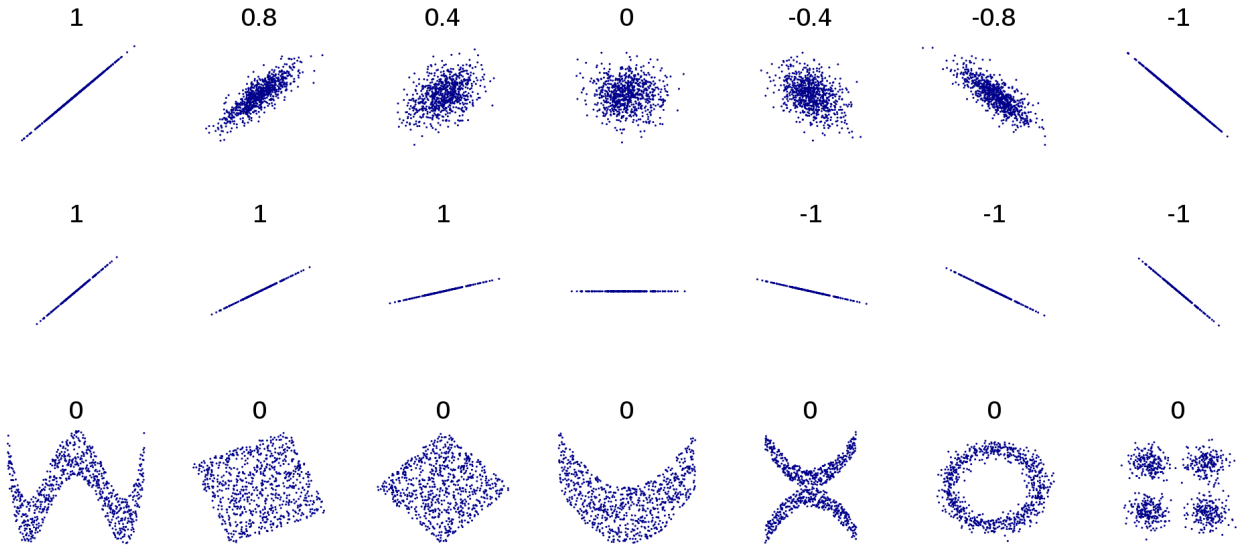


FIGURE 1.5 – Examples of point clouds with different correlation coefficients in a representation of Y as a function of X .

figure extracted from this page Wikipédia.

Attention : It is essential to understand that **correlation does not imply causation**. This confusion frequently appears in the media.

Let us consider all middle school students in France. Let X be the random variable representing their shoe size, and Y the random variable representing their level in mathematics. We

can all agree that a ninth-grade student (troisième) generally has a better level in mathematics than when they were in sixth grade (sixième), and we can also agree that their shoe size is likely to be larger in ninth grade than it was in sixth grade. Therefore, if we plot the point cloud of the points (x_i, y_i) for each student i, we would observe a strong correlation. However, there is, of course, no causal link between shoe size and mathematical ability.

Let us now consider another example involving countries. Each country has a certain number of Nobel Prize laureates among its citizens, and a corresponding annual chocolate consumption. One might again observe a strong correlation between the number of Nobel Prizes and chocolate consumption. Let's think for a moment : it is easy to understand that the higher a country's GDP (Gross Domestic Product), the more Nobel Prizes it will generate, because investing money in research increases the likelihood of winning Nobel Prizes. Likewise, the higher a country's GDP, the greater its consumption will be. Thus, it is the GDP that drives the increase of these two random variables. However, there is no causal link between chocolate consumption and the number of Nobel Prizes in a country.

Therefore, one must be extremely careful with this type of correlation and must never conclude a causal relationship. **Establishing causation is difficult in science**. We will come back to this when we study statistical tests.

1.4.2 The Linear Regression

Let us now try to fit a line in the "best possible way" through a point cloud.

Définition 1.17 (Regression line).
Let a sample of size n be drawn from two random variables X and Y, such that the variance σ_x^2 is non-zero. The **regression line of the variables X and Y** is the line defined by :

$$y = ax + b \text{ with } a = \frac{\sigma_{xy}}{\sigma_x^2} \text{ and } b = \overline{y} - a\overline{x}.$$

Proposition 1.6

This line passes through the point G and minimizes the **quadratic error** committed by the line :

$$E(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

This is known as the **least squares method**.

Démonstration. Let us calculate the minimum of this function E. It is a function of class $\mathcal{C}^1$ on $\mathbb{R}^2$, so at the minimum, its partial derivatives will be zero.

$$\left\{\begin{array}{l}\frac{\partial E}{\partial a}=0\\ \frac{\partial E}{\partial b}=0\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\sum_{i=1}^{n}-x_i\left(y_i-(ax_i+b)\right)=0\\ \sum_{i=1}^{n}-\left(y_i-(ax_i+b)\right)=0\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\sum_{i=1}^{n}\left(-x_iy_i+ax_i^2+bx_i\right)=0\\ \sum_{i=1}^{n}\left(-y_i+ax_i+b\right)=0\end{array}\right.$$
$$\Leftrightarrow \left\{\begin{array}{l}-\sum_{i=1}^{n}x_iy_i+a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=0\\ -\sum_{i=1}^{n}y_i+a\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}b=0\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}-\overline{xy}+a\overline{x^2}+b\overline{x}=0\\ -\overline{y}+a\overline{x}+b=0\end{array}\right.$$
$$\Leftrightarrow \left\{\begin{array}{l}-\overline{xy}+a\overline{x^2}+(\overline{y}-a\overline{x})\overline{x}=0\\ b=\overline{y}-a\overline{x}\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}a=\frac{\overline{xy}-\overline{x}\,\overline{y}}{\overline{x}^2-\overline{x}^2}=\frac{\sigma_{xy}}{\sigma_x^2}\\ b=\overline{y}-a\overline{x}\end{array}\right.$$

$\overline{y}=a\overline{x}+b$ proves that the center of gravity belongs to the regression line.

□

Exemple: Consider the following bivariate data set : $(-4,-1)$, $(-1,-1)$, $(2,2)$, and $(4,1)$. Let us plot the point cloud and the regression line :

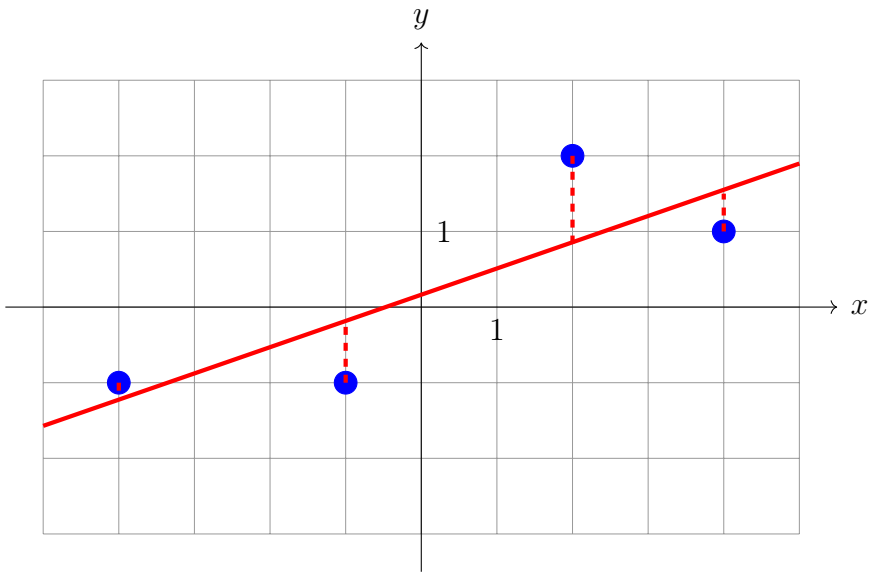


FIGURE 1.6 – The point cloud is shown in blue and the regression line in red. The dashed segments represent the errors made.

REMARK : The error committed could be defined differently : for example, one could evaluate the horizontal segments instead of the vertical segments. Another way to do this is to evaluate the segments orthogonal to the regression line.

1.5 Statistics on p variables

Let us generalize to an arbitrary number of variables.

1.5.1 Definitions

Consider, for example, the entire set of your $p = 80$ grades from the first semester for $n = 40$ students in your class. We will organize these values into a table. Each individual i, for $i \in [\![1, n]\!]$, can be represented by the p values obtained for each variable : $\begin{pmatrix} x_{i,1} & \cdots & x_{i,p} \end{pmatrix}$

For a given variable j, we thus have the n values of the individuals :

$$\begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$$

The concatenation forms the matrix representation of the raw data from which the analysis will be performed. It is denoted by M and has the following form :

$$\overbrace{\quad\quad\quad\quad\quad}^{p \text{ variables}}$$
$$n \text{ individuals} \left\{ \begin{pmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ \vdots & & \vdots & & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,p} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,p} \end{pmatrix} \right. = (x_{ij})_{ij} = M$$

Exemple: Consider a class of $n = 3$ students, each having received 4 grades in the first semester. The first student received $\begin{pmatrix} 8 & 10 & 11 & 12 \end{pmatrix}$. The second student received $\begin{pmatrix} 9 & 13 & 13 & 7 \end{pmatrix}$. The third student received $\begin{pmatrix} 14 & 14 & 14 & 17 \end{pmatrix}$. The concatenation is $M = \begin{pmatrix} 8 & 10 & 11 & 12 \\ 9 & 13 & 13 & 7 \\ 14 & 14 & 14 & 17 \end{pmatrix}$.

Définition 1.18 (Data Matrices).
Let a sample of n individuals and p random variables be given, denoted by X_j for $j \in [\![1, p]\!]$. We call :

- **Centered Matrix** : $M_c = (x_{i,j} - \overline{x_j})_{i,j}$
- **Centered-Reduced Matrix** : $M_s = \left(\frac{x_{i,j} - \overline{x_j}}{\sigma_j}\right)_{i,j}$

with $\overline{x_j}$ the average and σ_j the observed standard deviation of the variable X_j in the sample.

Définition 1.19 (Matrix of Analysis).
Consider a sample of n individuals and consider p random variables, denoted X_j for $j \in [\![1, p]\!]$. We call :

- **Variance-Covariance Matrix** : $\Sigma = \frac{1}{n}{}^t M_c . M_c$
- **Correlation Matrix** : $R = \frac{1}{n}{}^t M_s . M_s$

with $\overline{x_j}$ the average and σ_j the observed standard deviation of the variable X_j in the sample. The transpose of M is denoted ${}^t M$.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_{p-1,p} \\ \sigma_{p,1} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{pmatrix}_{p\times p} \quad , \quad R = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,p} \\ \rho_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{p-1,p} \\ \rho_{p,1} & \cdots & \rho_{p,p-1} & 1 \end{pmatrix}_{p\times p}$$

Proposition 1.7

The variance-covariance matrix and the correlation matrix are two symmetric $p \times p$ matrices.

For Σ : the variances are on the diagonal and the covariances are off the diagonal.

For R : ones are on the diagonal and the correlations are off the diagonal.

1.5.2 The tools to manipulate the data

Définition 1.20 (Euclidean Distance).
Let a sample of n individuals and p random variables be given, denoted X_j for $j \in [\![1, p]\!]$. The **distance between individuals** i **and** k is defined as :

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

Exemple: In the previous example $X = \begin{pmatrix} 8 & 10 & 11 & 12 \\ 9 & 13 & 13 & 7 \\ 14 & 14 & 14 & 17 \end{pmatrix}$

The distance between the students 1 and 3 is :

$d(x_1, x_3) = \sqrt{(8-14)^2 + (10-14)^2 + (11-14)^2 + (12-17)^2} = \sqrt{36 + 16 + 9 + 25} = \sqrt{86}$

1.6 Exercises

* = direct application of the course (should be solvable without any help) ;
** = more difficult (but should be solvable with some guidance if needed) ;
*** = challenging (bonus exercise for the most motivated students).

Exercice 1 (*).

Consider four students, each characterized by two values representing their academic year. The first value corresponds to a random variable called X_1, and the second to a variable called X_2. The four students have the following (x_1, x_2) values :

$$(-2,-2),\ (-3,-1),\ (2,3),\ (2,4)$$

1) Draw the point cloud associated with the data.
2) Write the matrix M of the values.
3) Compute the sample average and sample standard deviation of variable X_1. Do the same for variable X_2.
4) Compute the skewness coefficient S_k and the kurtosis K of variable X_1 in the sample. Do the same for variable X_2.
5) Compute the median, mode, and range of variable X_1 in the sample. Do the same for variable X_2.
6) Compute the correlation coefficient between the two variables.
7) Compute the least squares regression line.
8) Provide the centered matrix X_c, then the centered-reduced matrix X_s.
9) Compute, using matrix multiplication, the variance-covariance matrix Σ and the correlation matrix R.

   Identify each term and verify the previously computed results.

Exercice 2 (*).

In one room, 9 people are seated, and their average age is 25 years. In another room, 11 people are gathered, and their average age is 45 years. The two groups are then combined. Compute the average age of the resulting group.

Exercice 3 (*).

In a room, several families are gathered. It turns out that 25% of the families have one child, 40% have two children, 20% have three children, and 15% have four children. What is the average number of children per family in this group ?

Exercice 4 (*).

In a company with 360 employees, the executives earn an average of €2000 and the workers earn an average of €1200. The overall average salary for all employees is €1400.

1) How many executives are there ?
2) If all the workers receive a 5% raise :
   - **i)** What is the new average salary of the company's employees ?
   - **ii)** Does the range of salaries increase ?
   - **iii)** Does the median salary increase ?
3) Answer the same questions if, instead of increasing the workers' salaries, all the executives receive a 5% raise.

Exercice 5 (*).

In response to the question, « Statistics allow one to lie with confidence : what is your opinion ? », 80 people were surveyed and responded as follows :

Strongly disagree	10
Somewhat disagree	15
Neutral	12
Somewhat agree	18
Strongly agree	25

Let X be the variable associated with this survey.

1) What type of data does X provide ? What are the categories (modalities) of X ? What is its mode ?
2) Establish the frequency distribution of this variable and represent it using a bar chart.
3) What is the proportion of subjects who do not have an extremely strong opinion on the question ? What is the proportion of subjects who have an extremely strong opinion ?
4) What is the proportion of subjects who respond negatively to the question ? What is the proportion of subjects who respond positively to the question ?

Exercice 6 (**).

A vehicle is propelled at very high speed along an axis, then it slows down. We are interested in the speed of this vehicle during braking. Throughout the exercise, distances are expressed in meters, time in seconds, and thus speeds in meters per second. Results will be rounded to the nearest tenth. The instantaneous speeds v_i of this vehicle were recorded at times t_i, for i ranging from 0 to 7.

1) Plot the graph of this statistical series and explain why an affine (linear) adjustment of this point cloud is not considered.

t_i in s	0	1	2	3	4	5	6	7
v_i in $m.s^{-1}$	215	140	85	57	36	29	27	22

2) Define $n_i = \ln(v_i - 15)$ for i ranging from 0 to 7. Construct the table of the series (t_i, n_i).

3) Give an equation of the regression line of n as a function of t using the least squares method.

4) Deduce an expression of the speed v as a function of time t in the form $v = \alpha e^{\beta t} + \gamma$, where α, β and γ are real numbers to be determined, and plot the resulting curve on the graph from question 1.

Exercice 7 (***).

Consider the statistical series with three variables as follows : $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, and $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Find the regression plane by drawing inspiration from the two-variable case studied in the course.