# TP10 : Statistiques descriptives univariées

L'objectif de ce TP est de présenter le vocabulaire élémentaire de l'analyse statistique de données numériques ainsi que les commandes informatiques usuelles permettant d'obtenir les valeurs des indicateurs de répartition, de tendance centrale et de dispersion associés aux séries statistiques quantitatives.

## 1 Vocabulaire élémentaire de la statistique descriptive

**Définition 1.1 : Population, individus, échantillon**

Une analyse statistique consiste à synthétiser, organiser, résumer, afin de mieux les exploiter, des informations relatives à une **population**, modélisée par un ensemble fini ou infini, que l'on notera $\Omega$.

Les éléments de cette population sont appelés les **individus**, traditionnellement notés $\omega$.

Un **échantillon** extrait de cette population est la donnée d'un entier naturel $n \in \mathbb{N}^*$ et d'un $n$-uplet $(\omega_1, \dots, \omega_n) \in \Omega^n$ dont les composantes sont des individus deux à deux distincts de la population.

L'entier naturel $n$ s'appelle alors la **taille** de l'échantillon.

En pratique, on a rarement accès à l'opportunité d'observer tous les individus d'une population donnée, et on travaille uniquement sur les données récoltées sur un échantillon de la population, qui constitue donc la partie **observée** de la population dans l'étude statistique considérée.

Les données récoltées sur l'échantillon observé de la population peuvent être de natures diverses et multiples : on peut récolter une base de données contenant de nombreuses informations pour chaque individu et l'analyse de telles données constitue le domaine des statistiques multivariées. En ECG1, on se limitera à l'étude d'une seule information collectée pour chaque individu de l'échantillon, c'est-à-dire au cadre des **statistiques descriptives univariées**. En ECG2, les concepts seront étendus à l'analyse de deux informations par individu de l'échantillon, c'est-à-dire au cadre des **statistiques descriptives bivariées**. Le contexte général des statistiques multivariées soulève des questions complexes dont certaines alimentent des domaines de la recherche scientifique actuelle.

**Définition 1.2 : Caractères, variables statistiques**

Une étude statistique univariée porte sur un **caractère** de la population.

On distingue deux types de caractères que sont :

- les caractères qualitatifs tels que la couleur des yeux, la nationalité, le métier exercé, le sport favori, ...
- les caractères quantitatifs dont les **valeurs** sont **numériques** tels que l'âge, la taille, le nombre d'enfants à charge, le revenu mensuel net, ...

Un caractère s'appelle également **variable statistique** et est formellement la donnée d'une application $X$ définie sur $\Omega$. Dans le cas d'une variables statistique quantitative, il s'agit plus précisément d'une application définie sur $\Omega$ et à valeurs dans $\mathbb{R}$.

L'**ensemble image** $X(\Omega)$ d'une variable statistique forme l'ensemble des valeurs prises par $X$.

Dans toute la suite de ce TP, nous nous restreindrons au cas des variables statistiques **quantitatives** prenant un nombre fini de valeurs, c'est-à-dire telles que $X(\Omega)$ est un ensemble fini, que l'on appelle variables statistiques **discrètes**.

Dans ce cadre, on se propose de définir un certain nombre d'indicateurs permettant de dégager des caractéristiques interprétables du caractère étudié sur l'échantillon observé.

**Définition 1.3 : Modalités**

Les valeurs prises par une variable statistique discrète $X$ sont appelées les **modalités** de $X$.

Dans toute la suite, on notera $p \in \mathbb{N}^*$ le nombre de modalités observées sur l'échantillon étudié et on notera $x_1, \dots, x_p$ ces modalités avec la convention que $x_1 < \dots < x_p$ (c'est-à-dire que les modalités sont numérotées par ordre croissant de valeur numérique).

Autrement dit, on a $X(\Omega) = \{x_1, \dots, x_p\}$ avec $x_1 < \dots < x_p$ et $p = \text{Card}(X(\Omega))$.

**Exercice 1** Une collecte de données pour l'observation d'un caractère quantitatif $X$ (ici l'âge en années) d'un échantillon d'une population est transmise sous forme de liste ci-dessous :

```
L = [18,20,19,22,13,18,16,21,16,21,18,22,19,24,20,21,16,22,19,18]
```

Déterminer la taille de l'échantillon observé ainsi que l'ensemble des modalités $X(\Omega)$.

Écrire le code d'une fonction Python, nommée `modalites`, prenant en paramètre une liste numérique `L` et renvoyant en sortie une liste obtenue à partir de `L` en supprimant les éventuelles répétitions, triée par ordre strictement croissant (on pourra relire rapidement les exercices 10 et 19 du TP04). Tester la fonction `modalites` en prenant en argument la liste `L` ci-dessus, pour retrouver les résultats établis à la main.

Le concept le plus élémentaire de l'analyse de données est la notion d'effectif d'une modalité, correspondant au nombre d'occurences de cette modalité dans l'échantillon observé, ainsi que la notion associée de fréquence d'une modalité permettant d'évaluer l'importance relative de cette modalité dans l'échantillon.

**Définition 1.4 : Effectif d'une modalité**

Pour tout $i \in [\![1,p]\!]$, on appelle **effectif** de la modalité $x_i \in X(\Omega)$, et on note $n_i$, le nombre entier naturel :

$$n_i = \text{Card}(\{j \in [\![1,n]\!], X(\omega_j) = x_i\})$$

Autrement dit, l'effectif de la modalité $x_i$ est égal au nombre d'individus de l'échantillon pour lesquels le caractère observé prend la valeur $x_i$. Il s'agit d'une notion **qui dépend de l'échantillon considéré**.

La taille $n$ de l'échantillon est aussi appelé **effectif total**. On a nécessairement l'égalité $\sum_{i=1}^{p} n_i = n$.

L'effectif d'une modalité dans un échantillon est une information brute qui ne peut pas être traitée de façon pertinente. Savoir que l'effectif d'une modalité est égal à 42 ne permet pas de donner une quelconque interprétation, en l'absence de la connaissance de la taille de l'échantillon. Dans un échantillon de taille 50 ou dans un échantillon de taille $10^{10}$, une modalité d'effectif 42 n'aura pas vraiment le même poids et ne sera pas interprétée de la même façon. Il est donc nettement plus pertinent de considérer les proportions d'un échantillon correspondant à chaque modalité, que l'on appelle les fréquences.

**Définition 1.5 : Fréquence d'une modalité**

Pour tout $i \in [\![1,p]\!]$, on appelle **fréquence** de la modalité $x_i \in X(\Omega)$, et on note $f_i$, le nombre réel :

$$f_i = \frac{n_i}{n}$$

Observons que pour tout $i \in [\![1,p]\!]$, $f_i \in [0,1]$ et on a nécessairement l'égalité $\sum_{i=1}^{p} f_i = \sum_{i=1}^{p} \frac{n_i}{n} = \frac{n}{n} = 1$.

**Exercice 2** Déterminer l'effectif et la fréquence de chaque modalité de la liste de l'exercice 1.

**Définition 1.6 : Série statistique brute, série statistique dépouillée**

On appelle **série statistique brute** la donnée du $n$-uplet de réels $(X(\omega_1), \dots, X(\omega_n)) \in \mathbb{R}^n$.

On appelle **série statistique dépouillée** la donnée du $p$-uplet de couples $((x_1,n_1), \dots, (x_p,n_p)) \in (\mathbb{R}^2)^p$.

Une série statistique brute correspond à une remontée sans traitement des données collectées sur le terrain. Avant toute analyse statistique, il convient de procéder au dépouillement de la série qui implique d'identifier les modalités et de les trier par ordre croissant, puis de déterminer l'effectif de chaque modalité. L'analyse s'effectue ensuite en travaillant sur la série statistique dépouillée obtenue.

**Exercice 3** Écrire le code d'une fonction Python, nommée `depouille`, prenant en paramètre une liste numérique `L` contenant les valeurs d'une série statistique brute et renvoyant en sortie la série statistique dépouillée sous forme de liste de listes.

Par exemple, si `L = [20,12,33,12,33,12]`, alors la commande `print(depouille(L))` devra afficher en réponse la liste de listes `[[12,3],[20,1],[33,2]]`.

Tester la fonction `depouille` en prenant en argument la liste de l'exercice 1.

Afin de définir certains indicateurs de position associées à une série statistique, il sera également intéressant de manipuler les effectifs ou les fréquences sous formes cumulées.

**Définition 1.7 : Effectifs cumulés, fréquences cumulées**

Pour tout $i \in [\![1,p]\!]$, on appelle **effectif cumulé** de la modalité $x_i \in X(\Omega)$, et on note $n_i^c$, le nombre entier naturel :

$$n_i^c = \text{Card}\{j \in [\![1,n]\!], X(\omega_j) \leq x_i\}$$

Pour tout $i \in [\![1,p]\!]$, on appelle **fréquence cumulée** de la modalité $x_i \in X(\Omega)$, et on note $f_i^c$, le nombre réel :

$$f_i^c = \frac{n_i^c}{n}$$

**Proposition 1.8 : Expressions sommatoires des effectifs cumulés et des fréquences cumulées**

Pour tout $i \in [\![1,p]\!]$, $n_i^c = \sum_{k=1}^{i} n_k$. De plus, $n_1 = n_1^c \leq n_2^c \leq \cdots \leq n_p^c = n$.

Pour tout $i \in [\![1,p]\!]$, $f_i^c = \sum_{k=1}^{i} f_k$. De plus, $f_1 = f_1^c \leq f_2^c \leq \cdots \leq f_p^c = 1$.

*Démonstration.*

□

**Exercice 4** Déterminer l'effectif cumulé et la fréquence cumulée de chaque modalité de la liste de l'exercice 1.

**Exercice 5** Écrire le code d'une fonction Python, nommée `cumul`, prenant argument une liste `L` représentant une série statistique dépouillée au format `[[x1,n1],...,[xp,np]]` et renvoyant en sortie la liste donnant les effectifs cumulés de chaque modalité au format `[[x1,nc1],...,[xp,ncp]]`.

Tester cette fonction pour la série statistique dépouillée associée à l'exercice 1 en prenant en argument la liste obtenue en résultat de l'exerice 3.

**Exercice 6** Écrire le code d'une fonction Python, nommée `cumul_inverse`, prenant argument une liste `L` donnant les effectifs cumulés de chaque modalité au format `[[x1,nc1],...,[xp,ncp]]` et renvoyant en sortie une liste donnant la série statistique dépouillée au format `[[x1,n1],...,[xp,np]]`.

Tester cette fonction pour la série statistique dépouillée associée à l'exercice 1 en prenant en argument la liste obtenue en résultat de l'exercice 5.

Une série statistique dépouillée est souvent représentée sous forme de tableau :

| Modalités | $x_1$ | ... | $x_p$ |
|---|---|---|---|
| Effectifs | $n_1$ | ... | $n_p$ |

En Python, on peut transformer une liste de listes donnant une série statistique dépouillée au format `L = [[x1,n1],...,[xp,np]]` en tableau avec la syntaxe `np.array(np.transpose(L))` (en ayant au préalable importé le module `numpy` avec l'alias `np`).

```
import numpy as np

L = [[15,4], [17,3], [18,2], [19,6], [20,5], [21,5]]
A = np.array(np.transpose(L))
print(A)
```

Certaines séries statistiques contiennent des données numériques obtenues à partir de mesures réelles effectuées avec une certaine précision et pouvant prendre des valeurs non entières. Il est très fréquent pour ce type de collecte de données d'obtenir des valeurs observées toutes distinctes. Dans un tel contexte, la notion de fréquence d'une modalité perd tout son intérêt puisque tous les effectifs sont égaux à 1. Il est courant, pour pallier ce défaut, de recourir dans ce cas à des regroupements en classes, permettant ensuite de raisonner sur des catégories de modalités.

**Définition 1.9 : Regroupement en classes**

Soit $X$ une variable statistique quantitative.

Un **regroupement en classes** des modalités de $X$ est la donnée d'un entier naturel $q \in \mathbb{N}^*$ et de $q$ intervalles deux à deux disjoints de la forme :

$$I_1 = [y_1, y_2[, I_2 = [y_2, y_3[, \dots, I_q = [y_q, y_{q+1}[$$

tels que $X(\Omega) \subset [y_1, y_{q+1}[$.

Pour tout $i \in [\![1, q]\!]$, on dit que l'intervalle $I_i = [y_i, y_{i+1}[$ est une **classe** du caractère étudié $X$.

 Dans le cas d'un regroupement en classes $[y_1, y_2[$, $[y_2, y_3[$, $\dots$, $[y_q, y_{q+1}[$, les valeurs $y_1, \dots, y_{q+1}$ ne sont pas nécessairement des modalités $x_1, \dots, x_p$ (même si ce n'est pas interdit).

On s'intéresse alors aux effectifs et aux fréquences de chacune des classes pour mener une analyse statistique de l'échantillon observé.

**Définition 1.10 : Effectif d'une classe**

Pour tout $i \in [\![1, q]\!]$, on appelle **effectif** de la classe $I_i = [y_i, y_{i+1}[$, et on note $n_i$, le nombre entier naturel :

$$n_i = \text{Card}(\{j \in [\![1, n]\!], X(\omega_j) \in [y_i, y_{i+1}[\})$$

Autrement dit, l'effectif de la classe $[y_i, y_{i+1}[$ est égal au nombre d'individus de l'échantillon pour lesquels le caractère observé prend une valeur appartenant à cette classe.

Il s'agit d'une notion **qui dépend de l'échantillon considéré**.

**Définition 1.11 : Fréquence d'une classe**

Pour tout $i \in [\![1, q]\!]$, on appelle **fréquence** de la classe $I_i = [y_i, y_{i+1}[$, et on note $f_i$, le nombre réel :

$$f_i = \frac{n_i}{n}$$

Dans ce contexte de regroupement en classes, on peut également définir les notions d'effectifs cumulés et de fréquences cumulées de chaque classe.

**Définition 1.12 : Effectif cumulé d'une classe, fréquence cumulée d'une classe**

Pour tout $i \in [\![1, q]\!]$, on appelle **effectif cumulé** de la classe $I_i = [y_i, y_{i+1}[$, et on note $n_i^c$, le nombre entier naturel :

$$n_i^c = \text{Card}\{j \in [\![1, n]\!], X(\omega_j) \in [y_1, y_{i+1}[\}$$

Pour tout $i \in [\![1, q]\!]$, on appelle **fréquence cumulée** de la classe $I_i = [y_i, y_{i+1}[$, et on note $f_i^c$, le nombre réel :

$$f_i^c = \frac{n_i^c}{n}$$

**Exercice 7** Pour les données de la liste de l'exercice 1, on choisit d'effectuer un regroupement en classes avec les intervalles $[12, 15[$, $[15, 18[$, $[18, 21[$, $[21, 24[$ et $[24, 27[$.

Déterminer les effectifs, effectifs cumulés, fréquences et fréquences cumulées de chaque classe.

Dans toute démarche d'analyse d'une série statistique quantitative, qu'elle soit directe à partir des modalités ou que l'on utilise un regroupement en classes, il est essentiel d'utiliser des outils de visualisation des données. Les diagrammes bâtons sont adaptés pour la représentation des effectifs des modalités, tandis que les histogrammes sont adaptés pour la représentation des fréquences des classes.

**Définition 1.13 : Diagramme bâton (modalités)**

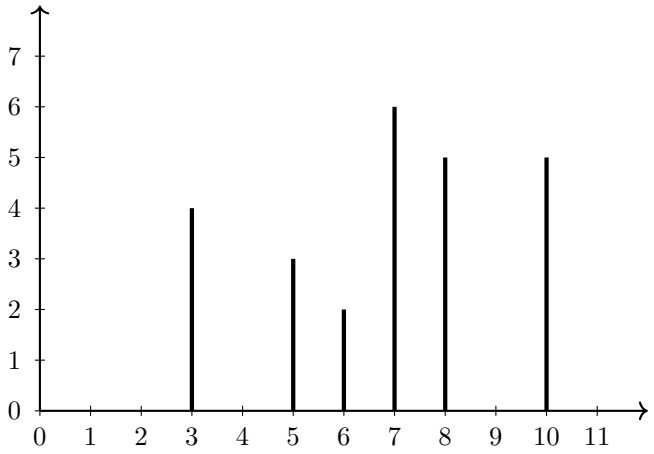
On considère une série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$.

Le **diagramme bâton** associé à cette série est l'union dans le plan muni d'un repère cartésien usuel des $p$ segments d'extrémités $(x_1, 0)$ et $(x_1, n_1)$, puis $(x_2, 0)$ et $(x_2, n_2)$, …, et enfin $(x_p, 0)$ et $(x_p, n_p)$.

En pratique, les segments sont tracés avec une certaine épaisseur par commodité de lecture, c'est pourquoi on les appelle des bâtons. Dans une telle représentation, chaque bâton a une hauteur égale à l'effectif de la modalité qu'il représente, et donc proportionnelle à la fréquence de cette modalité.

On représente ci-dessous le diagramme bâton associé à la série statistique dépouillée :

| Modalités | 3 | 5 | 6 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|
| Effectifs | 4 | 3 | 2 | 6 | 5 | 5 |




**Définition 1.14 : Histogramme normalisé (regroupement en classes)**

On considère une série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$.

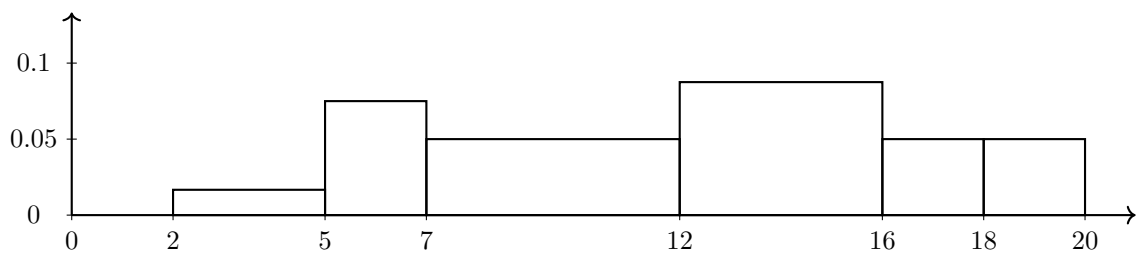
On considère un regroupement en classes $[y_1, y_2[$, $[y_2, y_3[$, …, $[y_q, y_{q+1}[$.

L'**histogramme normalisé** associé à ce regroupement en classes est l'union dans le plan muni d'un repère cartésien usuel des $q$ rectangles dont les bases sont les classes $[y_1, y_2[$, $[y_2, y_3[$ … , $[y_q, y_{q+1}[$ et de hauteurs respectives $\dfrac{f_1}{y_2 - y_1}$, $\dfrac{f_2}{y_3 - y_2}$, …, $\dfrac{f_q}{y_{q+1} - y_q}$ (et donc d'aires respectives $f_1$, $f_2$, …, $f_q$).

Par construction, dans un histogramme normalisé, la somme des aires des rectangles vaut $f_1 + f_2 + \cdots + f_q = 1$.

On représente ci-dessous le diagramme bâton associé au regroupement en classes :

| Classes | $[2, 5[$ | $[5, 7[$ | $[7, 12[$ | $[12, 16[$ | $[16, 18[$ | $[18, 20[$ |
|---|---|---|---|---|---|---|
| Fréquences | $\frac{1}{20}$ | $\frac{3}{20}$ | $\frac{5}{20}$ | $\frac{7}{20}$ | $\frac{2}{20}$ | $\frac{2}{20}$ |



En Python, les représentations graphiques de diagrammes bâton et histogrammes normalisés sont accessibles via le module `matplotlib.pyplot`, dont on rappelle qu'il doit être importé avec l'alias `plt`, grâce aux fonctions `plt.bar` et `plt.hist` respectivement.

**Définition 1.15 : Diagrammes bâton en Python**

On considère une série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$.

On suppose que l'on dispose d'une liste `M = [x1,...,xp]` contenant les modalités.

On suppose que l'on dispose d'une liste `E = [n1,...,np]` contenant les effectifs.

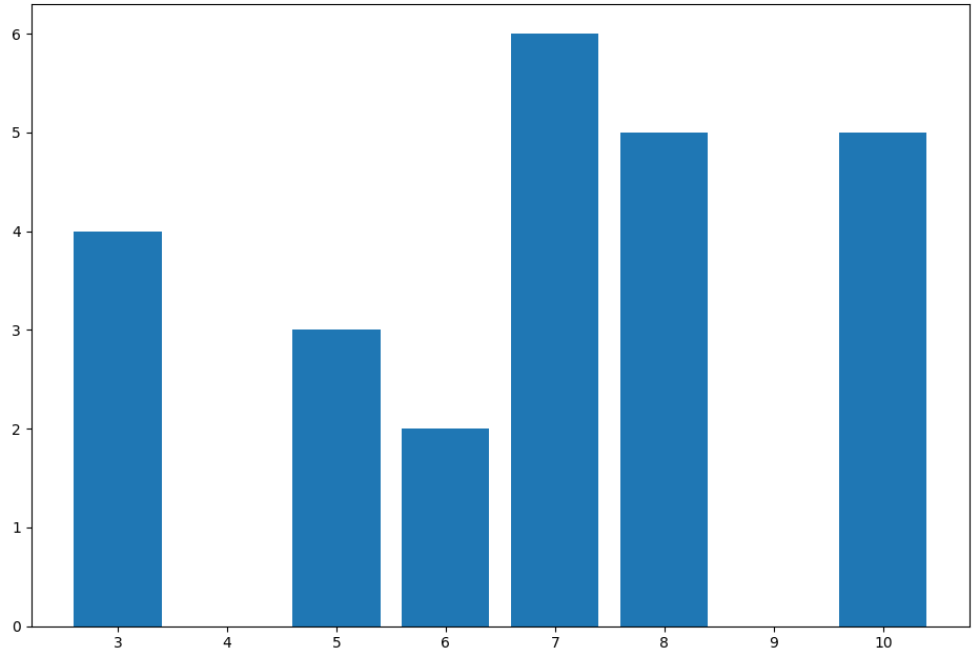
Alors pour obtenir le diagramme bâton associée à cette série statistique, on utilise les commandes :

```
plt.bar(M,E)
plt.show()
```

```
import matplotlib.pyplot as plt

M = [3,5,6,7,8,10]
E = [4,3,2,6,5,5]

plt.bar(M,E)
plt.show()
```




**Définition 1.16 : Histogrammes normalisés en Python**

On considère une série statistique brute $(X(\omega_1), \dots, X(\omega_n))$.

On considère un regroupement en classes $[y_1, y_2[, \dots, [y_q, y_{q+1}[$.

On suppose que l'on dispose d'une liste `L = [a1,...,an]` contenant la série statistique brute.

On suppose que l'on dispose d'une liste `C = [y1,...,yq,y(q+1)]` contenant les extrémités des classes.

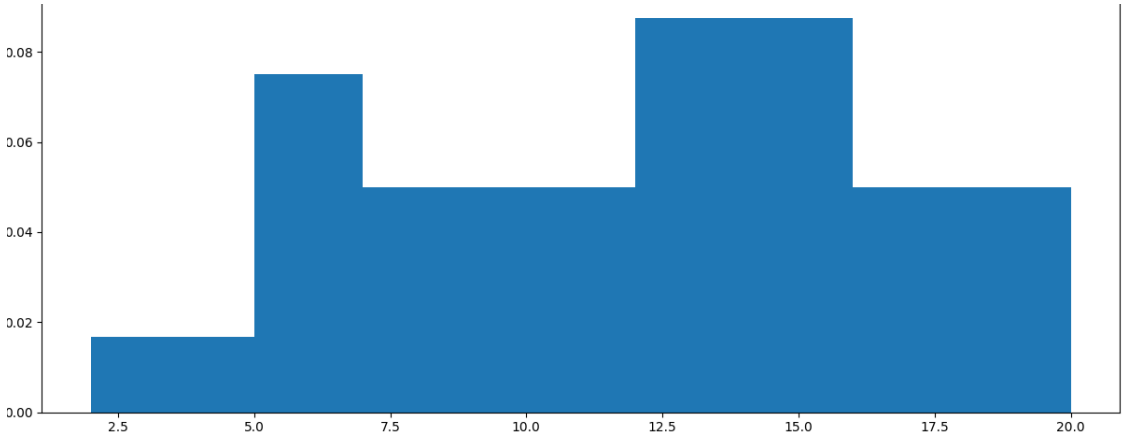
Alors pour obtenir l'histogramme normalisé associée à cette série statistique, on utilise les commandes :

```
plt.hist(L,C,density=True)
plt.show()
```

```
import matplotlib.pyplot as plt

L = [3,5,6,6,7,9,10,10,11,12,13,13,13,14,15,15,17,17,18,19]
C = [2,5,7,12,16,18,20]

plt.hist(L,C,density=True)
plt.show()
```



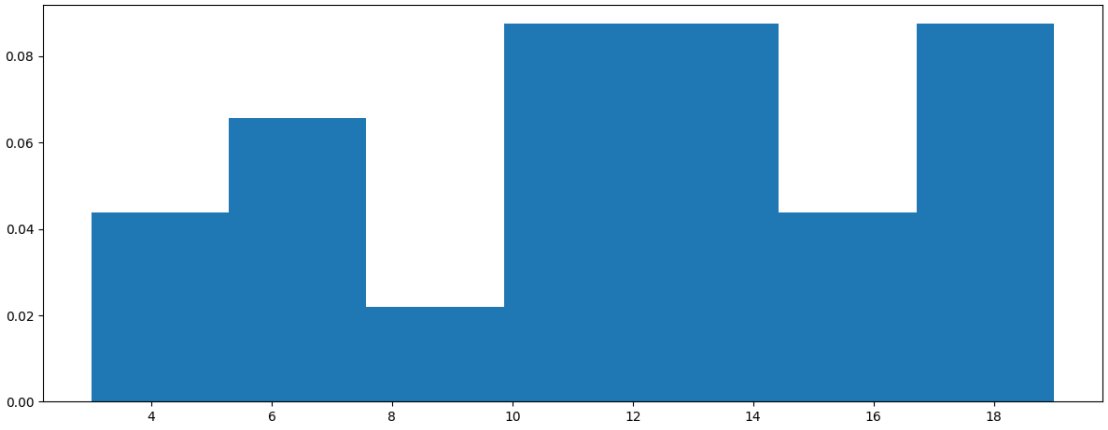

Il n'est pas toujours nécessaire de définir à la main les extrémités des classes dans une liste `C`. Il est également possible d'utiliser la fonction `plt.hist` en se contentant de choisir le **nombre** $q$ de classes que l'on souhaite utiliser.

Dans ce cas, l'histogramme sera construit automatiquement avec $q$ classes de même longueur allant de la plus petite à la plus grande des valeurs observées dans la série statistique.

```
import matplotlib.pyplot as plt

L = [3,5,6,6,7,9,10,10,11,12,13,13,13,14,15,15,17,17,18,19]
q = 7

plt.hist(L,q,density=True)
plt.show()
```




**Exercice 8** Écrire un code Python permettant d'afficher le diagramme bâton associé à la série statistique de l'exercice 1.

**Exercice 9** Écrire un code Python permettant d'afficher l'histogramme normalisé associé à la série statistique de l'exercice 1, en utilisant le regroupement en classes défini dans l'exercice 7.

**Exercice 10** Écrire un code Python permettant d'afficher l'histogramme normalisé associé à la série statistique de l'exercice 1, en utilisant un regroupement en 10 classes de même longueur.

Le dernier outil graphique utilisé pour les analyses statistiques est le diagramme des fréquences cumulées.

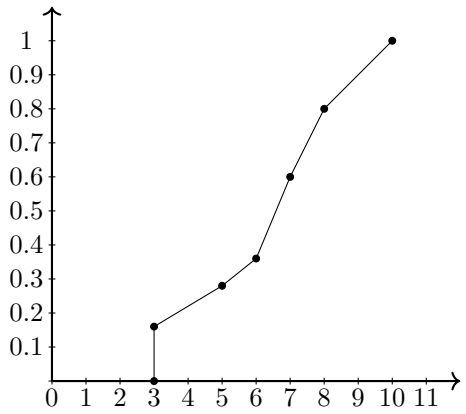
**Définition 1.17 : Diagramme des fréquences cumulées (modalités)**

On considère une série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$.

Le diagramme des fréquences cumulées est la courbe brisée dans le plan muni d'un repère cartésien usuel reliant consécutivement les points de coordonnées $(x_1, 0)$, $(x_1, f_1^c)$, $(x_2, f_2^c)$, $\dots$, $(x_p, f_p^c)$.

On représente ci-dessous le diagramme des fréquences cumulées associé à la série statistique dépouillée :

| Modalités | 3 | 5 | 6 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|
| Effectifs | 4 | 3 | 2 | 6 | 5 | 5 |




**Définition 1.18 : Diagramme des fréquences cumulées (regroupement en classes)**

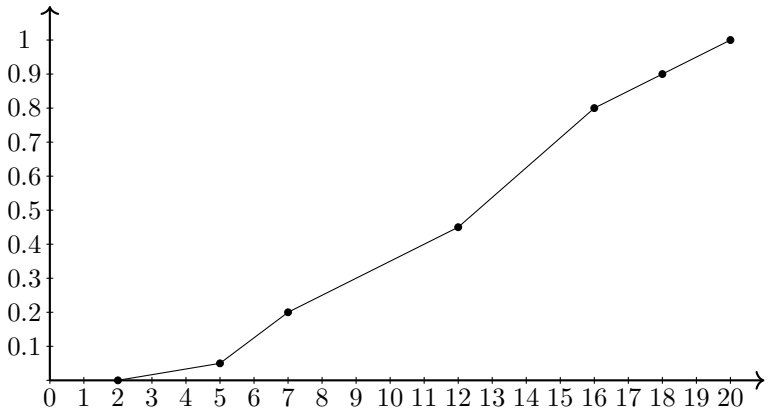
On considère une série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$.

On considère un regroupement en classes $[y_1, y_2[$, $[y_2, y_3[$, $\dots$, $[y_q, y_{q+1}[$.

Le diagramme des fréquences cumulées est la courbe brisée dans le plan muni d'un repère cartésien usuel reliant consécutivement les points de coordonnées $(y_1, 0)$, $(y_2, f_1^c)$, $(y_3, f_2^c)$, $\dots$, $(y_q, f_{q-1}^c)$, $(y_{q+1}, f_q^c)$.

On représente ci-dessous le diagramme des fréquences cumulées associé au regroupement en classes :

| Classes | $[2, 5[$ | $[5, 7[$ | $[7, 12[$ | $[12, 16[$ | $[16, 18[$ | $[18, 20[$ |
|---|---|---|---|---|---|---|
| Fréquences | $\frac{1}{20}$ | $\frac{3}{20}$ | $\frac{5}{20}$ | $\frac{7}{20}$ | $\frac{2}{20}$ | $\frac{2}{20}$ |




**Exercice 11** Écrire un code Python permettant d'afficher le diagramme des fréquences cumulées associé à la série statistique de l'exercice 1.

**Exercice 12** Écrire un code Python permettant d'afficher le diagramme des fréquences cumulées associé à la série statistique de l'exercice 1, en utilisant le regroupement en classes défini dans l'exercice 7.

## 2 Traitement des données avec le module numpy

Le traitement des données statistiques brutes apparaît comme un élément essentiel si l'on veut pouvoir être en mesure de représenter, synthétiser, interpréter les informations collectées. On présente ci-dessous les commandes du module numpy permettant de réaliser le plus simplement possible ces opérations courantes.

**Définition 2.1 : Conversions (fonction np.array)**

Si L est une liste numérique contenant les données d'une série statistique brute, alors la commande :

```
np.array(L)
```

convertit cette liste en vecteur numpy.

```
L = [2,5,1,3,2,2,5,1,2,1]

A = np.array(L)
print(A)
```

**Définition 2.2 : Modalités (fonction np.unique)**

Soit A un vecteur contenant les données d'une série statistique brute. Alors la commande :

```
M = np.unique(A)
```

définit un vecteur M contenant les modalités ordonnées par ordre croissant de la série statistique.

```
A = np.array([2,5,1,3,2,2,5,1,2,1])

M = np.unique(A)
print(M)
```

**Définition 2.3 : Modalités et effectifs (fonction np.unique, paramètre return_counts)**

Soit A un vecteur numpy contenant les données d'une série statistique brute. Alors la commande :

```
M, E = np.unique(A, return_counts=True)
```

définit deux vecteurs M et E, le premier contenant les modalités ordonnées par ordre croissant de la série statistique et le second contenant les effectifs correspondants.

La commande F = E/len(A) permet de définir un vecteur F contenant les fréquences correspondantes.

```
A = np.array([2,5,1,3,2,2,5,1,2,1])

M, E = np.unique(A, return_counts=True)
F = E/len(A)
print(M)
print(E)
print(F)
```

Ainsi, si A est un vecteur numpy contenant les données d'une série statistique brute, on peut obtenir une représentation graphique de ces données sous forme de diagramme bâton avec le code :

```
M, E = np.unique(A, return_counts=True)
plt.bar(M,E)
plt.show()
```

**Définition 2.4 : Sommes cumulées (fonction `np.cumsum`)**

Soit `A` un vecteur `numpy` à valeurs numériques. Alors la commande :

`np.cumsum(A)`

renvoit le vecteur des sommes cumulées des valeurs présentes dans `A`.

```
A = np.array([4,3,2,6,5,5])

B = np.cumsum(A)
print(B)
```

Ainsi, si `A` est un vecteur `numpy` contenant les données d'une série statistique brute, alors on peut obtenir une représentation du diagramme des fréquences cumulées avec le code :

```
M, E = np.unique(A, return_counts=True)
F = E/len(A)

FC = np.cumsum(F)

plt.plot(M,FC,'.-')
plt.show()
```

Dans le cas où l'on souhaite regrouper les modalités en classes, la construction du diagramme des fréquences cumulées est un peu plus complexe, mais réalisable en filtrant les données appartenant à chaque classe à l'aide d'une définition de liste en compréhension pour pouvoir déterminer leurs fréquences.

Si `A` est une liste ou un vecteur `numpy` contenant les données d'une série statistique brute et `C` est une liste ou un vecteur `numpy` contenant les extrémités des classes, alors on peut obtenir une représentation du diagramme des fréquences cumulées avec le code :

```
FC = [0]*len(C)
for i in range(1,len(C)):
  filtre = [A[k] for k in range(0,len(A)) if A[k]>=C[i-1] and A[k]<C[i]]
  FC[i] = len(filtre)/len(A) + FC[i-1]

plt.plot(C,FC,'.-')
plt.show()
```

La fonction `np.unique` n'apparaît pas dans le programme officiel. Elle pourra être utilisée librement dans le cadre de ce TP, mais il faut savoir proposer un code effectuant le même travail que cette commande, ce qui a été vu dans les exercices 1 et 3 à travers les fonctions `modalites` et `depouille`.

**Exercice 13** On considère la série statistique brute ci-dessous, donnant les résultats obtenus par une promotion d'étudiants en ECG à un devoir surveillé de Philosophie (imaginaire).

```
[9,7,8,11,10,9,8,7,10,7,6,10,9,8,6,9,7,7,13,8,7,8,9,8,7,14,15,7,8,9,7,9,8,6,9,7,9,8,7,7,10,9]
```

(*a*) Écrire un code Python affichant les modalités et effectifs de cette série statistique.

(*b*) Écrire un code Python affichant le diagramme bâton associé à cette série statistique.

(*c*) Écrire un code Python affichant le diagramme des fréquences cumulées associé à cette série statistique.

**Exercice 14** On considère la série statistique brute ci-dessous, donnant les résultats obtenus par une promotion d'étudiants en ECG à un devoir surveillé de Mathématiques (imaginaire).

```
[8,14.5,8.7,10,13.5,9.6,9.9,9.3,14,9.8,9.4,17.6,10.9,7.2,5.7,9.5,10.9,6.5,7.2,15.3,8.4,
9.7,16.3,8.6,9,17.5,9.4,4.5,9.7,9.7,8.1,6.7,16.7,6.7,5.9,12,8.5,8.8,8.1,18.3,8.7,10.5]
```

(*a*) Écrire un code Python affichant le diagramme des fréquences cumulées associé à cette série statistique.

(*b*) Écrire un code Python affichant le diagramme des fréquences cumulées associé à cette série statistique pour le regroupement en classes défini par la liste `C = [5,8,11,14,17,20]`

# 3 Indicateurs statistiques

Dans cette partie, on présente les indicateurs statistiques usuels permettant de résumer les informations contenues dans une série statistique, en distinguant les **indicateurs de position** permettant de situer les valeurs des **indicateurs de dispersion** permettant de mesurer la variabilité de ces valeurs.

Dans toute cette partie, on s'intéresse à l'observation d'une variable statistique quantitative discrète $X$ sur un échantillon $(\omega_1, \dots, \omega_n) \in \Omega^n$.

On notera, pour plus de commodité, $x = (v_1, \dots, v_n) = (X(\omega_1), \dots, X(\omega_n))$ la série statistique brute étudiée.

On conserve la notation $X(\Omega) = \{x_1, \dots, x_p\}$ avec $x_1 < \dots < x_p$ pour désigner les modalités ordonnées par ordre croissant, ainsi que les définitions des effectifs, effectifs cumulés, fréquences, fréquences cumulées, notés respectivement $n_i$, $n_i^c$, $f_i$ et $f_i^c$.

Les indicateurs de cette partie seront illustrés sur les séries statistiques brutes :

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 11 | 9 | 8 | 8 | 10 | 9 | 10 | 9 | 8 | 7 | 6 | 8 | 10 | 9 | 13 | 13 | 11 | 8 | 12 | 9 | 10 | 10 | 9 | 7 | 7 | 7 | 8 |

présentant un relevé des températures moyennes journalières à Paris en février 2022 (arrondies à l'unité) et :

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,37 | 1,10 | 1,32 | 1,46 | 1,54 | 1,52 | 1,51 | 1,29 | 1,26 | 1,42 | 1,47 | 1,41 | 1,55 | 1,40 | 1,65 |

présentant le prix moyen au litre en euros de l'essence SP95 au premier janvier pour les années 2008 à 2022.

## 3.1 Indicateurs de position

**Définition 3.1 : Mode**

Soit $i \in [\![1, p]\!]$. On dit que la modalité $x_i$ est un **mode** de la série $x$ si pour tout $j \in [\![1, p]\!]$, $n_i \geq n_j$.

**Exemple 3.2**

Pour les relevés de températures, on trouve le tableau des effectifs ci-dessous :

| $x_i$ | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 4 | 6 | 6 | 6 | 2 | 1 | 2 |

Donc la série statistique admet exactement 3 modes, donnés par les températures 8°C, 9°C et 10°C, toutes trois d'effectif maximal égal à 6.

Pour les relevés de prix d'essence, comme les valeurs sont sans répétitions, tous les effectifs sont égaux à 1, donc toutes les modalités sont des modes. Ainsi, la notion de mode n'a pas vraiment d'intérêt pour des séries dont les valeurs ne sont jamais répétées.

**Définition 3.3 : Minimum et maximum**

Soit $i \in [\![1, p]\!]$. On dit que la valeur $v_i$ est le **minimum** de la série $x$ si pour tout $j \in [\![1, p]\!]$, $v_i \leq v_j$.

Soit $i \in [\![1, p]\!]$. On dit que la valeur $v_i$ est le **maximum** de la série $x$ si pour tout $j \in [\![1, p]\!]$, $v_i \geq v_j$.

Ainsi, le minimum de la série est égal à la modalité $x_1$ et le maximum de la série est égal à la modalité $x_p$.

**Exemple 3.4**

Pour les relevés de température, le minimum de la série est égal à 6°C, température moyenne observée une unique fois le 12 février, tandis que le maximum de la série est égal à 13°C, température moyenne observée deux fois les 16 et 17 février.

Pour les prix d'essence, le minimum de la série est égal à 1,10€/litre, prix moyen observé au premier janvier 2009, tandis que le maximum de la série est égal à 1,65€/litre, prix moyen observé au premier janvier 2022. On peut observer que le minimum et le maximum peuvent beaucoup dépendre de la présence d'une valeur exceptionnelle, si l'on regarde uniquement la décennie entre 2010 et 2019, alors le minimum vaut 1,26€/litre et le maximum vaut 1,54€/litre, ce qui donne l'impression d'une moindre variabilité.

**Définition 3.5 : Moyenne (empirique) $\overline{x}$**

La **moyenne** de la série $x = (v_1, \dots, v_n)$, notée $\overline{x}$, est le nombre réel défini par $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} v_i$.

**Exemple 3.6**

Pour les prix de l'essence, on trouve :

$$\overline{x} = \frac{1,37+1,1+1,32+1,46+1,54+1,52+1,51+1,29+1,26+1,42+1,47+1,41+1,55+1,4+1,65}{15} \simeq 1,42$$

**Proposition 3.7 : Expression de la moyenne en fonction des effectifs ou des fréquences**

On a les égalités : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{p} x_i\, n_i = \sum_{i=1}^{p} x_i\, f_i$.

*Démonstration.*

□

**Remarque 3.8** Ainsi, la moyenne (arithmétique) des valeurs est précisément égale à la moyenne des modalités, pondérées par leurs fréquences. On pourra faire le lien entre cette notion est la notion d'espérance d'une variable aléatoire discrète dans le cours de probabilités.

**Exemple 3.9**

Pour les relevés de température, on peut utiliser le tableau des effectifs pour calculer la moyenne :

$$\overline{x} = \frac{6\times 1 + 7\times 4 + 8\times 6 + 9\times 6 + 10\times 6 + 11\times 2 + 12\times 1 + 13\times 2}{28} = \frac{64}{7} \simeq 9,14$$

**Définition 3.10 : Médiane $M_e$**

La **médiane** de la série $x = (v_1, \dots, v_n)$, notée $M_e$, est la plus petite modalité pour laquelle la fréquence cumulée dépasse $\dfrac{1}{2}$. Autrement dit, $M_e = x_{\min\left\{i\in[\![1,p]\!], f_i^c \geq \frac{1}{2}\right\}}$.

**Exemple 3.11**

Pour les relevés de températures, on trouve le tableau des fréquences cumulées ci-dessous :

| $x_i$ | 6 | 7 | 8 | **9** | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|
| $f_i^c$ | 0.036 | 0.180 | 0.394 | **0.608** | 0.822 | 0.893 | 0.929 | 1 |

Donc la médiane de la série est égale à $M_e = 9$°C.

Pour les prix de l'essence, on trouve le tableau de fréquences cumulées ci-dessous :

| $x_i$ | 1,1 | 1,26 | 1,29 | 1,32 | 1,37 | 1,4 | 1,41 | **1,42** | 1,46 | 1,47 | 1,51 | 1,52 | 1,54 | 1,55 | 1,65 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_i^c$ | 0.07 | 0.13 | 0.2 | 0.27 | 0.33 | 0.4 | 0.47 | **0.53** | 0.6 | 0.67 | 0.73 | 0.8 | 0.87 | 0.93 | 1 |

Donc la médiane de la série est égale à $M_e = 1,42$€/litre.

**Proposition 3.12 : Partage des valeurs en deux**

La médiane est une modalité qui partage la série statistique en deux au sens suivant :

$$\text{Card}(\{j \in [\![1,n]\!], v_j \leq M_e\}) \geq \frac{n}{2} \text{ et } \text{Card}(\{j \in [\![1,n]\!], v_j \geq M_e\}) \geq \frac{n}{2}$$

*Démonstration.*

□

Il est très facile de déterminer la médiane d'une série statistique brute dans le cas où les valeurs sont ordonnées par ordre croissant : dans ce cas, il suffit de prendre la valeur située au milieu de la liste si l'effectif total est impair ou juste avant le milieu de la liste si l'effectif total est pair.

**Définition 3.13 : Série statistique brute ordonnée**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute.

Notons $(v'_1, \dots, v'_n)$ le $n$-uplet formé des mêmes valeurs que la série $x$ tel que $v'_1 \leq \dots \leq v'_n$.

On dit alors que $x' = (v'_1, \dots, v'_n)$ est la **série statistique ordonnée** associée à $x$.

**Proposition 3.14 : Calcul de la médiane pour une série statistique ordonnée**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute et $x' = (v'_1, \dots, v'_n)$ la série ordonnée associée à $x$.

Alors la médiane $M_e$ de $x$ est donnée par $M_e = v'_{\lfloor \frac{n+1}{2} \rfloor} = \begin{cases} v'_{\frac{n+1}{2}} & \text{si } n \text{ est impair} \\ v'_{\frac{n}{2}} & \text{si } n \text{ est pair} \end{cases}$

*Démonstration.*

□

**Exemple 3.15**

Pour les relevés de températures, la série ordonnée $x' = (v'_1, \dots, v'_{28})$ est donnée par :

| $v'_i$ | 6 | 7 | 7 | 7 | 7 | 8 | 8 | 8 | 8 | 8 | 8 | 9 | 9 | **9** | 9 | 9 | 9 | 10 | 10 | 10 | 10 | 10 | 10 | 11 | 11 | 12 | 13 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Donc la médiane de la série est égale à $M_e = v'_{14} = 9$.

Pour les prix de l'essence, la série ordonnée $x' = (v'_1, \dots, v'_{15})$ est donnée par :

| $v'_i$ | 1,1 | 1,26 | 1,29 | 1,32 | 1,37 | 1,4 | 1,41 | **1,42** | 1,46 | 1,47 | 1,51 | 1,52 | 1,54 | 1,55 | 1,65 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Donc la médiane de la série est égale à $M_e = v'_8 = 1,42$.

**Remarque 3.16** On trouve parfois une définition alternative dans le cas où l'effectif total $n$ est pair. Dans ce cas, on peut appeler médiane tout nombre appartenant à l'intervalle $\left]v'_{\frac{n}{2}}, v'_{\frac{n}{2}+1}\right[$ que l'on appelle l'intervalle médian et le choix le plus souvent retenu est de prendre le milieu de cet intervalle pour définir $M_e = \frac{v'_{\frac{n}{2}} + v'_{\frac{n}{2}+1}}{2}$.

Avec une telle définition, si $v'_{\frac{n}{2}} \neq v'_{\frac{n}{2}+1}$, alors la médiane n'est pas une modalité de la série.

**Définition 3.17 : Premier quartile $q_1$, Troisième quartile $q_3$**

Le **premier quartile** de la série $x = (v_1, \dots, v_n)$, noté $q_1$, est la plus petite modalité pour laquelle la fréquence cumulée dépasse $\frac{1}{4}$. Autrement dit, $q_1 = x_{\min\left\{i \in [\![1,p]\!], f_i^c \geq \frac{1}{4}\right\}}$.

Le **troisième quartile** de la série $x = (v_1, \dots, v_n)$, noté $q_3$, est la plus petite modalité pour laquelle la fréquence cumulée dépasse $\frac{3}{4}$. Autrement dit, $q_3 = x_{\min\left\{i \in [\![1,p]\!], f_i^c \geq \frac{3}{4}\right\}}$.

**Exemple 3.18**

Pour les relevés de températures, on trouve le tableau des fréquences cumulées ci-dessous :

| $x_i$ | 6 | 7 | **8** | 9 | **10** | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|
| $f_i^c$ | 0.036 | 0.180 | **0.394** | 0.608 | **0.822** | 0.893 | 0.929 | 1 |

Donc le premier quartile de la série est égal à $q_1 = 8$°C.

Et le troisième quartile de la série est égal à $q_3 = 10$°C.

Pour les prix de l'essence, on trouve le tableau de fréquences cumulées ci-dessous :

| $x_i$ | 1,1 | 1,26 | 1,29 | **1,32** | 1,37 | 1,4 | 1,41 | 1,42 | 1,46 | 1,47 | 1,51 | **1,52** | 1,54 | 1,55 | 1,65 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_i^c$ | 0.07 | 0.13 | 0.2 | **0.27** | 0.33 | 0.4 | 0.47 | 0.53 | 0.6 | 0.67 | 0.73 | **0.8** | 0.87 | 0.93 | 1 |

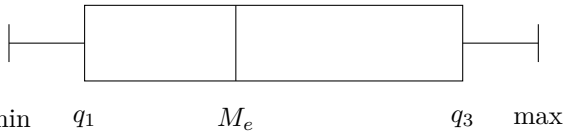
Donc le premier quartile de la série est égal à $q_1 = 1,32$€/litre.

Et le troisième quartile de la série est égal à $q_3 = 1,52$€/litre.

Les indicateurs de position que sont le minimum, le premier quartile, la médiane, le second quartile et le maximum se représentent souvent sous forme d'un diagramme inventé par l'américain Tukey en 1977 que l'on appelle communément "boîte à moustache" ("Box-and-whiskers plot", abrégé en box plot en anglais).

**Définition 3.19 : Boîte à moustache (box plot)**

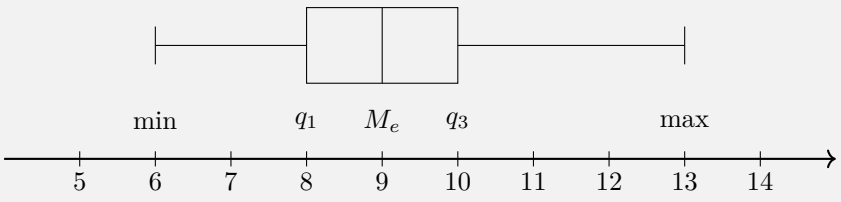
Soit $x = (v_1, \dots, v_n)$ une série statistique. Une représentation de la série sous forme de **boîte à moustaches** est formée de deux segments horizontaux allant respectivement du minimum au premier quartile et du troisième quartile au maximum, ainsi que d'un rectangle allant du premier au troisième quartile, séparé en deux verticalement au niveau de la médiane.




**Exemple 3.20**

Pour les relevés de températures, on a min = 6, $q_1 = 8$, $M_e = 9$, $q_3 = 10$ et max = 13.

Donc la représentation sous forme de boîte à moustaches et donnée par :




Une telle représentation permet de se faire une idée de la répartition des valeurs de la série statistique, relativement aux seuils correspondant aux proportions 0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ et 1 de l'effectif total.

## 3.2 Indicateurs de dispersion

**Définition 3.21 : Étendue $x_p - x_1$**

L'**étendue** de la série $x = (v_1, \dots, v_n)$ est la différence entre la modalité de plus grande valeur et la modalité de plus petite valeur. Autrement dit, l'étendue de $x$ est égale à $x_p - x_1$.

**Exemple 3.22**

Pour les relevés de températures, on a $p = 8$, $x_1 = 6$ et $x_8 = 13$.

Donc l'étendue de la série est égale à $x_8 - x_1 = 13 - 6 = 7$.

Pour les prix de l'essence, on a $p = 15$, $x_1 = 1,10$ et $x_{15} = 1,65$.

Donc l'étendue de la série est égale à $x_{15} - x_1 = 1,65 - 1,10 = 0,55$.

L'étendue est une mesure de dispersion très sensible aux valeurs extrêmes, pouvant être des aberrations dans la série statistique (venant par exemple d'erreurs de mesures). Il est parfois plus pertinent de supprimer les valeurs les plus grandes et les plus petites pour obtenir une mesure de la dispersion qui soit plus exploitable.

**Définition 3.23 : Écart inter-quartile $q_3 - q_1$**

L'**écart inter-quartile** de la série $x = (v_1, \dots, v_n)$ est la différence entre le troisième quartile et le premier quartile. Autrement dit, l'écart inter-quartile de $x$ est égal à $q_3 - q_1$.

L'intervalle $[q_1, q_3]$ est appelé l'**intervalle inter-quartile**.

**Exemple 3.24**

Pour les relevés de températures, on a $q_1 = 8$ et $q_3 = 10$.

Donc l'écart-inter-quartile de la série est égal à $q_3 - q_1 = 10 - 8 = 2$.

L'intervalle inter-quartile est l'intervalle $[8, 10]$ et son effectif est égal à 18.

Pour les prix de l'essence, on a $q_1 = 1,32$ et $q_3 = 1,52$.

Donc l'écart inter-quartile de la série est est égal à $q_3 - q_1 = 1,52 - 1,32 = 0,20$.

L'intervalle inter-quartile est l'intervalle $[1,32; 1,55]$ et son effectif est égal à 9.

**Proposition 3.25 : Effectif de l'intervalle inter-quartile**

On a l'inégalité : $\text{Card}(\{j \in [\![1, n]\!], v_j \in [q_1, q_3]\}) \geq \dfrac{n}{2}$.

*Démonstration.*

□

**Définition 3.26 : Variance (empirique) $s_x^2$**

La **variance** de la série $x = (v_1, \dots, v_n)$, notée $s_x^2$, est le nombre réel défini par $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(v_i - \overline{x})^2$.

La variance est donc la mesure moyenne des écarts quadratiques entre les valeurs de la série statistique et sa valeur moyenne. En particulier, on a toujours l'inégalité $s_x^2 \geq 0$. Plus la valeur de $s_x^2$ est grande, plus les valeurs observées ont tendance à être éloignées de $\overline{x}$, et inversement.

**Proposition 3.27 : Expression de la variance en fonction des effectifs ou des fréquences**

On a les égalités : $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{p}(x_i-\overline{x})^2\,n_i = \sum_{i=1}^{p}(x_i-\overline{x})^2\,f_i$.

*Démonstration.*

□

**Définition 3.28 : Écart-type (empirique) $\sigma_x$**

L'**écart-type** de la série $x=(v_1,\dots,v_n)$, noté $\sigma_x$, est le nombre réel défini par $\sigma_x = \sqrt{s_x^2}$.

**Proposition 3.29 : Formule de Kœnig-Huygens**

On a l'égalité : $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} v_i^2 - \overline{x}^2 = \dfrac{1}{n}\sum_{i=1}^{p} x_i^2\,n_i - \overline{x}^2 = \sum_{i=1}^{p} x_i^2\,f_i - \overline{x}^2$.

*Démonstration.*

□

En pratique, c'est le plus souvent avec la formule de Kœnig-Huygens que l'on calcule une variance.

**Exemple 3.30**

Pour les relevés de températures, on a :

- $\dfrac{1}{n}\sum_{i=1}^{p} x_i^2\,n_i = \dfrac{6^2\times 1+7^2\times 4+8^2\times 6+9^2\times 6+10^2\times 6+11^2\times 2+12^2\times 1+13^2\times 2}{28} = \dfrac{1213}{14}$
- $\overline{x}^2 = \left(\dfrac{64}{7}\right)^2 = \dfrac{4096}{49}$
- $s_x^2 = \dfrac{1213}{14} - \dfrac{4096}{49} = \dfrac{299}{98} \simeq 3,05$
- $\sigma_x = \sqrt{\dfrac{299}{98}} \simeq 1,75$

**Exemple 3.31**

Pour les prix de l'essence, on a :

- $\dfrac{1}{n}\sum_{i=1}^{p} v_i^2 = \dfrac{101\,417}{50\,000}$
- $\overline{x}^2 = \left(\dfrac{709}{500}\right)^2 = \dfrac{502\,681}{250\,000}$
- $s_x^2 = \dfrac{101\,417}{50\,000} - \dfrac{502\,681}{250\,000} = \dfrac{1101}{62500} \simeq 0,017616$
- $\sigma_x = \sqrt{\dfrac{1101}{62500}} = \dfrac{\sqrt{1101}}{250} \simeq 0,132725$

# 4 Fonction de répartition empirique et quantiles

Dans cette section, on introduit la notion de fonction de répartition empirique d'un échantillon, et celle de quantile qui généralise la définition des quartiles vue dans la section précédente.

**Définition 4.1 : Fonction de répartition empirique**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute.

La **fonction de répartition empirique** associée à cette série est la fonction :

$$\begin{array}{rccc} F_x : & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & \dfrac{\text{Card}(\{j \in [\![1,n]\!], v_j \leq t\})}{n} \end{array}$$

**Proposition 4.2 : Valeurs de la fonction de répartition empirique**

$$\text{Pour tout } t \in \mathbb{R},\ F_x(t) = \begin{cases} 0 & \text{si : } t \in ]-\infty, x_1[ \\ f_i^c & \text{si : } \exists i \in [\![1, p-1]\!], t \in [x_i, x_{i+1}[ \\ 1 & \text{si : } t \in [x_p, +\infty[ \end{cases}$$

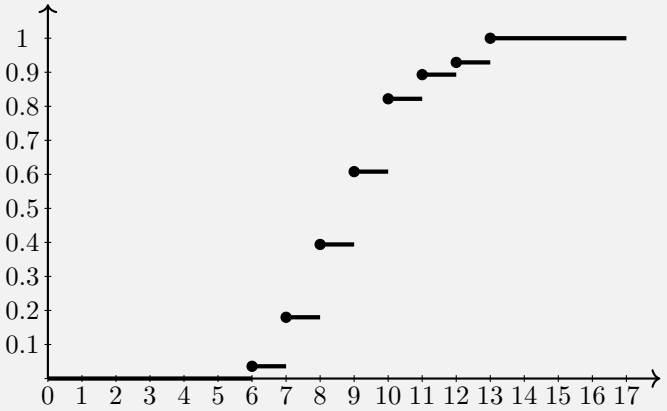
*Démonstration.*

□

Ainsi, la fonction de répartition empirique est la fonction constante par morceaux prenant sur chaque intervalle $[x_i, x_{i+1}[$ la valeur $f_i^c$ et égale à 0 sur $]-\infty, x_1[$ et à 1 sur $[x_p, +\infty[$. En particulier, il s'agit d'une fonction croissante sur $\mathbb{R}$, continue à droite mais pas à gauche en chaque modalité.

**Exemple 4.3**

Pour les relevés de températures, on obtient la fonction de répartition empirique ci-dessous :




**Définition 4.4 : Quantiles $Q_\alpha$**

Soit $\alpha \in ]0,1[$. Le **quantile d'ordre** $\alpha$ de la série $x = (v_1, \dots, v_n)$, noté $Q_\alpha$, est la plus petite modalité pour laquelle la fréquence cumulée dépasse $\alpha$. Autrement dit, $Q_\alpha = x_{\min\{i \in [\![1,p]\!], f_i^c \geq \alpha\}}$.

**Exemple 4.5**

La médiane et les quartiles sont des quantiles. On a précisément $M_e = Q_{\frac{1}{2}}$, $q_1 = Q_{\frac{1}{4}}$ et $q_3 = Q_{\frac{3}{4}}$.

**Définition 4.6 : Déciles $D_1, \dots, D_9$**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute.

Les **déciles**, notés, $D_1$, ..., $D_9$ de la série $x$ sont les quantiles d'ordre $\frac{1}{10}$, ..., $\frac{9}{10}$ respectivement. Autrement dit, $D_1 = Q_{\frac{1}{10}}$, ..., $D_9 = Q_{\frac{9}{10}}$.

L'**écart inter-décile** est la longueur $D_9 - D_1$ de l'**intervalle inter-décile** $[D_1, D_9]$ comprenant au moins 80% des effectifs de la série.

**Remarque 4.7** Le cinquième décile $D_5 = Q_{\frac{5}{10}} = Q_{\frac{1}{2}}$ n'est autre que la médiane $M_e$.

**Exemple 4.8**

Pour les relevés de températures, on rappelle le tableau des fréquences cumulées ci-dessous :

| $x_i$ | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|
| $f_i^c$ | 0.036 | 0.180 | 0.394 | 0.608 | 0.822 | 0.893 | 0.929 | 1 |

Donc les déciles de la série sont donnés par le tableau :

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | 7 | 8 | 8 | 9 | 9 | 9 | 10 | 10 | 12 |

L'intervalle inter-décile est l'intervalle $[7, 12]$ et l'écart inter-décile vaut $D_9 - D_1 = 5$.

Pour les prix de l'essence, on rappelle le tableau de fréquences cumulées ci-dessous :

| $x_i$ | 1,1 | 1,26 | 1,29 | 1,32 | 1,37 | 1,4 | 1,41 | 1,42 | 1,46 | 1,47 | 1,51 | 1,52 | 1,54 | 1,55 | 1,65 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_i^c$ | 0.07 | 0.13 | 0.2 | 0.27 | 0.33 | 0.4 | 0.47 | 0.53 | 0.6 | 0.67 | 0.73 | 0.8 | 0.87 | 0.93 | 1 |

Donc les déciles de la série sont donnés par le tableau :

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | 7 | 8 | 8 | 9 | 9 | 9 | 10 | 10 | 12 |

L'intervalle inter-décile est l'intervalle $[1,26; 1,55]$ et l'écart inter-décile vaut $D_9 - D_1 = 0,29$.

**Définition 4.9 : Centiles $C_1, \dots, C_{99}$**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute.

Les **centiles**, notés, $C_1$, ..., $C_{99}$ de la série $x$ sont les quantiles d'ordre $\frac{1}{100}$, ..., $\frac{99}{100}$ respectivement. Autrement dit, $C_1 = Q_{\frac{1}{100}}$, ..., $C_{99} = Q_{\frac{99}{10}}$.

L'**écart inter-centile** est la longueur $C_{99} - C_1$ de l'**intervalle inter-centile** $[C_1, C_{99}]$ comprenant au moins 98% des effectifs de la série.

**Remarque 4.10** Le cinquantième centile $C_{50} = Q_{\frac{50}{100}} = Q_{\frac{1}{2}}$ n'est autre que la médiane $M_e$.

**Exemple 4.11**

Pour les relevés de températures, on trouve $C_1 = 6$ et $C_{99} = 13$.

Donc l'écart inter-centile vaut $C_{99} - C_1 = 13 - 6 = 7$, ce qui dans ce cas est égal à l'étendue de la série.

Pour les prix de l'essence, on trouve $C_1 = 1,1$ et $C_{99} = 1,65$.

Donc l'écart inter-centile vaut $C_{99} - C_1 = 1,65 - 1,1 = 0,55$, ce qui dans ce cas est égal à l'étendue de la série.

Pour que l'utilisation des centiles et en particulier de l'intervalle inter-centile ait une utilité, il faut considérer des échantillons de plus grande taille, dépassant au minimum l'ordre de la centaine.

# 5 Analyse statistique avec le module `numpy`

**Exercice 15** Soit `A` un vecteur `numpy` contenant les données d'une série statistique brute quantitative.

La commande `np.sum(A)` renvoie la somme des valeurs du tableau `A`. On pourra utiliser cette commande (qui est officiellement au programme) pour répondre aux questions suivantes.

(*a*) Écrire un code Python permettant d'afficher les modes de la série `A`.
(*b*) Écrire un code Python permettant d'afficher le minimum de la série `A`.
(*c*) Écrire un code Python permettant d'afficher le maximum de la série `A`.
(*d*) Écrire un code Python permettant d'afficher la moyenne de la série `A`.
(*e*) Écrire un code Python permettant d'afficher la médiane de la série `A`.
(*f*) Écrire un code Python permettant d'afficher le premier quartile de la série `A`.
(*g*) Écrire un code Python permettant d'afficher le troisième quartile de la série `A`.
(*h*) Écrire un code Python permettant d'afficher l'étendue de la série `A`.
(*i*) Écrire un code Python permettant d'afficher l'écart inter-quartile de la série `A`.
(*j*) Écrire un code Python permettant d'afficher la variance de la série `A`.
(*k*) Écrire un code Python permettant d'afficher l'écart-type de la série `A`.

Tester chacun de ces codes pour les séries correspondant aux relevés de température et aux prix de l'essence présentées dans la section 3. Reprendre ensuite les mêmes questions en utilisant les commandes spécifiques du module `numpy` présentées ci-dessous, permettant d'obtenir plus directement certains indicateurs.

**Définition 5.1 : Indicateurs statistiques avec `numpy`**

Soit `A` un tableau `numpy` contenant des valeurs numériques.

Alors la commande :

- `np.min(A)` renvoie la valeur minimale de `A`
- `np.max(A)` renvoie la valeur maximale de `A`
- `np.mean(A)` renvoie la moyenne des valeurs de `A`
- `np.median(A)` renvoie la médiane de des valeurs de `A` (en choisissant la définition alternative où la médiane est égale au milieu de l'intervalle médian si `len(A)` est pair)
- `np.var(A)` renvoie la variance des valeurs de `A`
- `np.std(A)` renvoie l'écart-type des valeurs de `A`

Ces commandes sont toutes officiellement au programme.

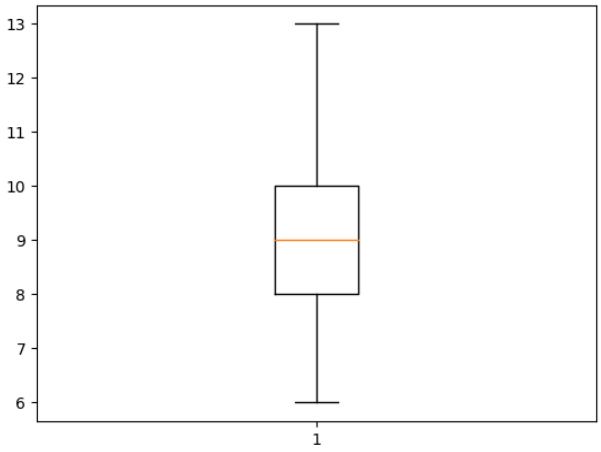
Enfin, on peut obtenir des représentations de séries statistiques sous forme de boîte à moustaches à l'aide de la fonction `boxplot` du module `matplotlib.pyplot`.

**Définition 5.2 : Représentation en boîte à moustaches (`boxplot`)**

Soit `A` un vecteur `numpy` contenant les valeurs d'une série statistique brute. Alors on peut obtenir une représentation graphique des valeurs de `A` sous forme de boîte à moustaches avec le code :

```
plt.boxplot(A,whis=[0,100])
plt.show()
```

Ci-contre, le diagramme obtenu pour les températures.



# 6 Exercices

**Exercice 16** On considère la série statistique quantitative brute $x$ ci-dessous :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $-1$ | 0 | $-2$ | 2 | $-1$ | 3 | 1 | 2 | 0 | 1 |

On répondra aux questions de cet exercice sans utiliser Python.

(*a*) Déterminer la série dépouillée associée à $x$, ainsi que le tableau des fréquences cumulées.

(*b*) Déterminer la moyenne $\overline{x}$ de $x$.

(*c*) Déterminer la variance $s_x^2$ et l'écart-type $\sigma_x$ de $x$.

(*d*) Déterminer la médiane $M_e$, le premier quartile $q_1$ et le troisième quartile $q_3$ de $x$.

(*e*) Déterminer l'étendue et l'écart inter-quartile de $x$.

(*f*) L'écart absolu moyen à la moyenne d'une série statistique brute $x = (v_1, \dots, v_n)$ est la valeur de $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} |v_i - \overline{x}|$.

Déterminer l'écart absolu moyen à la moyenne de la série étudiée dans cet exercice et comparer sa valeur à celle de $\sigma_x$. Comment interpréter cette comparaison ?

**Exercice 17** Déterminer à l'aide de Python les valeurs de tous les indicateurs statistiques mentionnés dans l'exercice 15, pour les séries statistiques des exercices 13 et 14.

**Exercice 18** Une entreprise fabrique des boîtes de clous contenant chacune environ 50 pièces.

Un prélèvement d'une vingtaine de boîtes est effectué au hasard. En comptant le nombre de clous de chaque boîte, on obtient la série statistique suivante :

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | 51 | 48 | 48 | 50 | 51 | 52 | 49 | 50 | 50 | 53 | 49 | 48 | 50 | 51 | 50 | 52 | 49 | 46 | 51 |

À l'aide de Python, déterminer les valeurs de tous les indicateurs statistiques mentionnés dans l'exercice 15 pour cette série. Représenter également ces données sous forme de diagramme des fréquences cumulées ainsi que sous forme de boîte à moustaches.

**Exercice 19** On considère la série statistique brute, contenant un échantillon de 50 mesures en centimètres de la taille d'individus humains, définie par la commande :

```
A = np.floor(np.random.randint(165,175)+np.random.randint(4,7)*np.random.randn(50))
```

À l'aide de Python, déterminer les valeurs de tous les indicateurs statistiques mentionnés dans l'exercice 15 pour cette série. Représenter également ces données sous forme de diagramme des fréquences cumulées ainsi que sous forme de boîte à moustaches.

**Exercice 20** On considère la série statistique brute, contenant un échantillon de 250 mesures en minutes du temps de trajet quotidien des salariés d'une grande entreprise, définie par la commande :

```
A = np.abs(np.floor(np.random.randint(25,35)+np.random.randint(20,25)*np.random.randn(250)))
```

À l'aide de Python, déterminer les valeurs de tous les indicateurs statistiques mentionnés dans l'exercice 15 pour cette série. Représenter également ces données sous forme de diagramme des fréquences cumulées ainsi que sous forme de boîte à moustaches.

Déterminer également l'intervalle inter-décile et l'intervalle inter-centile associés à cette série.

**Exercice 21** On considère la série statistique brute, contenant un échantillon de 100 salaires mensuels nets en euros d'employés d'une grande entreprise, définie par la commande :

```
A = 3000+np.abs(100*np.random.randn(100))-np.abs(500*np.random.randn(100))
```

On regroupe les données de cette série en classes $[1000, 1500[$, $[1500, 2000[$, $[2000, 2500[$, $[2500, 3000[$ et $[3000, 3500[$.

Dans une telle situation, on construit une série dépouillée $((x_1, n_1), \dots, (x_5, n_5))$ pour laquelle $x_1$ est le milieu de la classe $[y_1, y_2[$ et $n_1$ est son effectif, $\dots$, $x_5$ est le milieu de la classe $[y_5, y_6[$ et $n_5$ est son effectif.

On considère alors que les indicateurs statistiques de la série regroupée en classes sont ceux de cette série dépouillée. À l'aide de Python, déterminer la moyenne, la variance, la médiane et les quartiles de cette série.

**Exercice 22** Soit $x = (v_1, \dots, v_n)$ une série statistique. Pour tout $t \in \mathbb{R}$, on pose $f(t) = \displaystyle\sum_{i=1}^{p} (x_i - t)^2 f_i$.

Montrer que $f$ admet un unique extremum local sur $\mathbb{R}$ et qu'il s'agit d'un minimum global $m$ dont on déterminera la valeur ainsi que celle de son unique antécédent par $f$.

# 7 Aspects théoriques

On présente ci-dessous quelques propriétés théoriques des indicateurs statistiques usuels étudiés dans ce TP.

**Proposition 7.1 : Stabilité de l'opération de moyenne**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute et $((x_1, n_1), \dots, (x_p, n_p))$ la série dépouillée associée. Alors :

$$\overline{x} \in [x_1, x_p]$$

*Démonstration.*

□

**Proposition 7.2 : Échantillons de variance nulle**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute. Alors :

$$\sigma_x^2 = 0 \iff v_1 = \dots = v_n$$

*Démonstration.*

□

**Définition 7.3 : Transformation affine d'une série statistique**

Soit $x = (v_1, \dots, v_n)$ une série statistique quantitative brute, $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$.
On dit que la série statistique $y = ax + b = (av_1 + b, \dots, av_n + b)$ est une **transformée affine** de $x$.

**Proposition 7.4 : Linéarité de la moyenne**

Soit $x = (v_1, \dots, v_n)$ une série statistique quantitative brute, $a \in \mathbb{R}^*$, $b \in \mathbb{R}$ et $y = ax + b$. Alors :

$$\overline{y} = a\overline{x} + b$$

*Démonstration.*

□

**Définition 7.5 : Échantillons centrés**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute. On dit que $x$ est **centrée** si $\overline{x} = 0$.

**Proposition 7.6 : Mécanisme de recentrage**

Soit $x = (v_1, \dots, v_n)$ une série statistique brute. Alors $y = x - \overline{x}$ est centrée.

*Démonstration.*

□

**Proposition 7.7 : Propriété de la variance**

Soit $x = (v_1, \dots, v_n)$ une série statistique quantitative brute, $a \in \mathbb{R}^*$, $b \in \mathbb{R}$ et $y = ax + b$. Alors :

$$s_y^2 = a^2\, s_x^2$$

*Démonstration.*

□

**Proposition 7.8 : Propriété de l'écart-type**

Soit $x = (v_1, \dots, v_n)$ une série statistique quantitative brute, $a \in \mathbb{R}^*$, $b \in \mathbb{R}$ et $y = ax + b$. Alors :

$$\sigma_y = |a|\, \sigma_x$$

*Démonstration.*

□

**Exercice 23** On considère un relevé de températures $x$ effectué quotidiennement sur une semaine, les mesures ayant été effectuées en degrés Fahrenheit :

| | | | | | | |
|---|---|---|---|---|---|---|
| 77 | 68 | 69,8 | 71,6 | 73,4 | 78,8 | 82,4 |

(*a*) Déterminer les valeurs de $\overline{x}$ et $\sigma_x$.

(*b*) La température $t_F$ en degré Fahrenheit est liée à la température $t_C$ en degré Celsius par la relation :

$$t_F = \frac{9}{5}\, t_C + 32$$

Pour répondre aux questions suivantes, on ne convertira pas les températures mesurées.

(*i*) Déterminer la température moyenne observée sur la semaine, en degré Celsius.

(*ii*) Déterminer également l'écart-type des températures observées sur la semaine, en degré Celsius.

**Proposition 7.9 : Moyenne de la fusion de deux échantillons**

Soient $x = (v_1, \dots, v_n) \in \mathbb{R}^n$ et $y = (w_1, \dots, v_m) \in \mathbb{R}^m$ deux séries statistiques quantitatives brutes.

On pose $z = (v_1, \dots, v_n, w_1, \dots, w_m) \in \mathbb{R}^{n+m}$. Alors :

$$\overline{z} = \frac{n\,\overline{x} + m\,\overline{y}}{n+m}$$

*Démonstration.*

□

**Exercice 24** Le taux d'emploi (en pourcentage) de la population âgée de 25 à 49 ans en France de 2006 à 2021 est une série statistique $x$ donnée par le tableau ci-dessous :

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 81,8 | 82,6 | 83,7 | 82,3 | 82,1 | 81,6 | 80,9 | 80,6 | 80,4 | 80,0 | 80,3 | 80,7 | 81,0 | 81,5 | 81,1 | 81,8 |

Le taux d'emploi (en pourcentage) de la population âgée de 25 à 49 ans en France de 1994 à 2005 est une série statistique $y$ donnée par le tableau ci-dessous :

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 79,3 | 79,9 | 79,9 | 79,2 | 79,7 | 80,0 | 81,0 | 81,8 | 81,8 | 81,2 | 81,2 | 81,3 |

Déterminer les valeurs de $\overline{x}$ et $\overline{y}$ et en déduire la valeur du taux d'emploi moyen de la population agée de 25 à 49 ans en France de 1994 à 2021 en effectuant le minimum de calculs.