

TP de Python numéro 10 : Statistiques descriptives univariées

Semaine du 19 mars.

L'objectif de ce TP est de présenter le vocabulaire élémentaire de l'analyse statistique, les indicateurs de répartition classiques et quelques commandes Python liées aux problématiques du domaine.

I. Statistiques descriptives univariées

L'analyse statistique s'intéresse à l'organisation de données récoltées sur une population : après avoir collecté des informations auprès d'un certain nombre d'individus d'une population, on souhaite décrire la répartition des données récoltées afin d'en dégager une compréhension globale.

La population étudiée sera modélisée par un ensemble Ω . On appellera "individu" tout élément w de Ω .

En pratique, on ne "sonde" pas tous les individus d'une population lors d'une étude statistique : les individus "sondés" sont désignés par la notion d'échantillon.

Définition 1. Notons Ω un ensemble modélisant la population d'une étude statistique.

1. Les éléments ω de Ω sont appelés **les individus**.
2. On appelle **échantillon** extrait de la population Ω la donnée d'un entier $n \in \mathbb{N}^*$ et d'un n -uplet $(\omega_1, \dots, \omega_n) \in \Omega^n$ formé d'individus deux à deux distincts. L'entier n est appelé **la taille de l'échantillon**.

Les données récoltées sur un échantillon observé peuvent être de nature diverse.

Tout d'abord, on distingue le nombre d'information récoltées par individu.

- On parle de statistiques descriptives **univariées** lorsqu'on récolte une unique information par individu. On se placera dans ce contexte en ECG1.
- On parle de statistique descriptive **multivariées** pour des études statistiques dans lesquelles plusieurs informations sont récoltées pour chaque individu. En deuxième année, vous étudierez le cas des statistiques descriptives bivariées, où deux informations sont récoltées par individu.

La notion de **variable statistique** désignera l'information qu'on va récolter. Le terme *univariée* désigne alors le fait de ne s'intéresser qu'à une variable statistique.

Ensuite, on distingue la nature des informations récoltées. Ces informations peuvent être :

- ou bien des caractères **qualitatifs**, tels que la couleur des yeux, le métier exercé, ou le dessert favori,
- ou bien des caractères **quantitatifs**, dès que l'information récoltée est directement numérique : âge, taille, revenu mensuel net, nombre estimé de fois qu'on peut manger son dessert favori lors d'un repas...

Exemple 2. Considérons l'ensemble Ω formé par la population française en cet instant.

L'âge constitue un caractère quantitatif observable sur la population française. Pour chaque français ω , notons $A(\omega)$ l'âge de ω . On définit ainsi une application A , donnant l'âge :

$$A : \begin{cases} \Omega & \longrightarrow \mathbb{R} \\ \omega & \longmapsto A(\omega) \end{cases} .$$

La doyenne des français a, à ce jour, 114 ans : L'ensemble des âges possibles pour un individu de Ω est l'ensemble $A(\Omega) = \llbracket 0, 114 \rrbracket$.

Définition 3. 1. On appelle **variable statistique quantitative** sur une population Ω toute application

$$X : \Omega \rightarrow \mathbb{R}.$$

2. L'**ensemble image** $X(\Omega)$ d'une telle variable statistique quantitative forme donc l'ensemble des valeurs prises par X . Ces valeurs prises par X sont appelées les **modalités** de X .
3. Si l'ensemble $X(\Omega)$ des modalités est fini, on dit que X est une variable statistique quantitative **discrète**.

1. Notations pour la suite de ce TP

A partir de maintenant, afin d'alléger la suite de ce TP, on va se placer dans le cadre d'une étude statistique univariée et fixer quelques notations pour nos énoncés.

- On fixe une population, que l'on note Ω .
- On fixe une variable statistique discrète X sur Ω .
- On fixe un entier $n \in \mathbb{N}^*$ et un échantillon $(\omega_1, \dots, \omega_n) \in \Omega^n$.
- On note $p \in \mathbb{N}^*$ le nombre de modalités de X , et on note $x_1, \dots, x_p$ ces modalités (distinctes) triées dans l'ordre croissant ($x_1 < x_2 < \dots < x_p$). Autrement dit :

$$\begin{cases} p = \text{Card}(X(\Omega)) \\ X(\Omega) = \{x_1, \dots, x_p\} \\ x_1 < \dots < x_p \end{cases}$$

Exercice 4. Une collecte de données pour l'observation de l'âge des individus d'une population Ω est transmise via la liste suivante :

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

1. Déterminer à la main la taille de l'échantillon observé et les modalités observées.
2. Écrire une fonction `modalites` prenant en entrée une liste numérique L , représentant l'observation d'une variable statistique X sur un échantillon, et renvoyant en sortie la liste $[x_1, \dots, x_p]$ des modalités observées (triées dans l'ordre croissant).
3. Tester la fonction `modalites` sur la liste L ci-dessus et vérifier que l'on retrouve les résultats trouvés à la main.

2. Effectif et fréquence d'une modalité. Dépouillement d'une série statistique.

L'effectif d'une modalité x_i est simplement le nombre de fois que cette modalité a été observée sur notre échantillon $X(\omega_1), \dots, X(\omega_n)$. C'est donc le nombre d'individus observés dont le caractère décrit par X prend la valeur x_i . Cette quantité **dépend de l'échantillon considéré**.

Définition 5. Pour tout $i \in \llbracket 1, p \rrbracket$, on appelle **effectif de la modalité** $x_i \in X(\Omega)$ le nombre entier naturel, noté n_i dans toute la suite, donné par :

$$n_i = \text{Card}(\{k \in \llbracket 1, n \rrbracket \mid X(\omega_k) = x_i\}).$$

L'entier n (la taille de l'échantillon) est appelé **l'effectif total**.

Remarque. On a donc : $\sum_{i=1}^p n_i = n$.

L'effectif d'une modalité est donc le nombre de fois que celle-ci apparaît, mais c'est une information brute qui ne prend du sens que lorsque celle-ci est comparée à l'effectif total : observer 40 individus de 18 ans n'a pas du tout le même sens sur un échantillon de 100 individus que sur un échantillon de 10000 individus.

La notion de fréquence est plus directement pertinente.

Définition 6. Pour tout $i \in \llbracket 1, p \rrbracket$, on appelle **fréquence de la modalité** $x_i \in X(\Omega)$ le nombre réel, noté f_i dans toute la suite, donné par :

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

Remarque. $\sum_{i=1}^p f_i = \dots$

Exercice 7. Déterminer à la main les effectifs et fréquences des modalités observées dans la liste de l'exercice 4.

Pour présenter les résultats d'une étude statistique, on peut d'abord donner la liste des observations $(X(\omega_1), \dots, X(\omega_n))$ obtenues. On parlera de série statistique brute.

Pour rendre ceci plus facile à exploiter, on procède généralement au dépouillement de cette liste : on donne les modalités et les effectifs de chacune de celles-ci.

Définition 8. • On appelle **série statistique brute** la donnée du n -uplet $(X(\omega_1), \dots, X(\omega_n))$.
• On appelle **série statistique dépouillée** la donnée du p -uplet (formé de couples) :
 $((x_1, n_1), (x_2, n_2), \dots, (x_p, n_p))$.

Remarque. En informatique, une série statistique brute sera donnée par une liste, une série statistique dépouillée sera donnée par une liste de listes chacune de longueur 2.

Exercice 9. Ecrire une fonction `depouille` prenant en entrée une liste numérique représentant une série statistique brute, et renvoyant en sortie la série statistique dépouillée obtenue à partir de celle-ci.

Par exemple, si `L = [ 19, 20, 11, 20, 12, 19, 18 ]`, alors `depouille(L)` devra renvoyer la liste

$$[[11, 1], [12, 1], [18, 1], [19, 2], [20, 2]]$$

Tester la fonction `depouille` avec la liste de l'exercice 4.

3. Effectifs cumulés, fréquences cumulées

Définition 10. Soit $i \in \llbracket 1, p \rrbracket$.

- On appelle **effectif cumulé** de la modalité x_i le nombre entier, noté n_i^c , donné par :

$$n_i^c = \text{Card}(\{j \in \llbracket 1, n \rrbracket | X(\omega_j) \leq x_i\}).$$

- On appelle **fréquence cumulée** de la modalité x_i le réel noté f_i^c donné par :

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

Proposition 11. Pour tout $i \in \llbracket 1, p \rrbracket$:

- $n_i^c = \sum_{k=1}^i n_k$. De plus, $n_1 = n_1^c \leq n_2^c \leq \dots \leq n_p^c = n$.
- $f_i^c = \sum_{k=1}^i f_k$. De plus, $f_1 = f_1^c \leq f_2^c \leq \dots \leq f_p^c = 1$.

Démonstration. En exercice. $\square$

Exercice 12. 1. Déterminer à la main l'effectif cumulé et la fréquence cumulée de chaque modalité de la liste de l'exercice 4.

2. Écrire le code d'une fonction `cumul` prenant en entrée une liste `L` représentant une série statistique dépouillée (au format `[[x1, n1], ..., [xp, np]]`) et renvoyant en sortie la liste donnant les modalités et effectifs cumulés (au format `[[x1, n1c], ..., [xp, npc]]`).

3. Écrire le code d'une fonction `cumul_inverse` réalisant l'opération inverse, c'est-à-dire prenant en entrée une liste représentant des modalités et effectifs cumulés au format précédent, et renvoyant en sortie la série statistique dépouillée qui lui est associée.
4. Tester vos fonctions sur la liste de l'exercice 4.

4. Commandes Python liées aux notions précédentes

On supposera le module `numpy` importé avec la commande

```
import numpy as np
```

Alors, on manipule généralement les données sous la forme d'un vecteur `numpy`.

Conversion d'une liste en vecteur Numpy

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

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

converti `L` en vecteur `Numpy`.

Exemple 13.

```
L=[6,2,3,1]
print(L) # Affiche [6, 2, 3, 1]
print(type(L)) # Affiche <class 'list'>

L=np.array(L)
print(L) # Affiche [6 2 3 1]
print(type(L)) # Affiche <class 'numpy.ndarray'>
```

Modalités

Soit `L` un vecteur `Numpy` représentant les données d'une série statistique brute.

Alors, la commande

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

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

Exemple 14.

```
A = np.array([2,5,1,3,2,2,5,1,2,1])
M = np.unique(A)
print(M) # Affiche [1 2 3 5]
```

Remarque. Certaines commandes Python peuvent être appelées avec des paramètres supplémentaires, pour changer leurs effets. C'est le cas de la commande `np.unique`.

Modalités et effectifs

Soit `L` un vecteur `Numpy` représentant les données d'une série statistique brute.

Alors, la commande

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

définit :

- un vecteur `M` contenant les modalités ordonnées par ordre croissant de cette série statistique, et
- un vecteur `E` contenant les effectifs de ces modalités, dans leur ordre d'apparition.

Exemple 15. Prévoir l'affichage des commandes suivantes.

```
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) # Affiche ...
print(E) # Affiche ...
print(F) # Affiche ...
```

Remarque. Attention, la fonction `np.unique` n'apparaît pas dans le programme officiel : il est nécessaire de savoir refaire les fonctions `modalites` et `depouille` vues précédemment.

Pour avoir les effectifs cumulés, on peut utiliser la fonction `cumsum` de Numpy.

Le vecteur des sommes cumulées d'un vecteur numérique $[l_1, \dots, l_n]$ est simplement le vecteur numérique

$$[l_1, l_1 + l_2, l_1 + l_2 + l_3, \dots, l_1 + l_2 + \dots + l_n].$$

sommes cumulées d'un vecteur Numpy

Soit L un vecteur Numpy numérique.

Alors, la commande `np.cumsum(L)` renvoie le vecteur Numpy des sommes cumulées de L .

Exemple 16.

```
A=np.array([6,2,3,1,3])
print(np.cumsum(A)) # Affiche [6 8 11 12 15]
```

Exemple 17. Voici des choses dont on aura souvent besoin.

En présence d'une série statistique brute L ,

- On définira les vecteurs M et E des modalités et des effectifs avec la commande

$$M, E = np.unique(L, return_counts = True)$$

- On définira le vecteur des effectifs cumulés avec la commande

$$NC = np.cumsum(E)$$

- On définira le vecteur des fréquences avec la commande

$$F = E/len(L)$$

- On définira le vecteur des fréquences cumulées avec la commande

$$FC = NC/len(L)$$

ou avec la commande

$$FC = np.cumsum(F)$$

II. Représentations graphiques et séries statistiques

On va s'intéresser aux différents graphiques qu'on peut produire pour représenter les résultats d'une étude statistique.

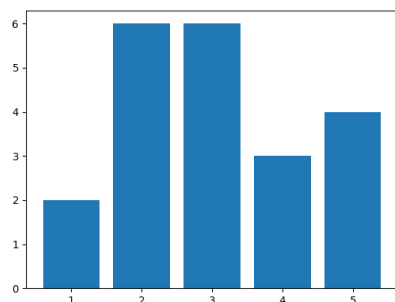
On aura besoin du module `matplotlib.pyplot`, que l'on supposera importé à l'aide de la commande

```
import matplotlib.pyplot as plt
```

1. Diagrammes en bâtons

Un **diagramme en bâtons** représente, de la manière la plus simple possible, une série statistique dépouillée obtenue : on place, dans un graphique, un segment vertical au dessus de chaque modalité dont la hauteur donne l'effectif de chaque modalité.

Par exemple, ce diagramme en bâtons représente la série statistique dont les modalités et effectifs sont donnés par le tableau ci-dessous.



Modalités	1	2	3	4	5
Effectifs	2	6	6	3	4

Diagramme en bâtons en Python

Soit L un vecteur Numpy contenant une série statistique brute.

Notons M le vecteur Numpy des modalités présentes dans L et E le vecteur Numpy donnant, dans le même ordre, les effectifs de ces modalités.

Alors :

- la commande `plt.bar(M,E)` génère le diagramme en bâton représentant cette série statistique,
- la commande `plt.show()` demande à Python d'afficher ce graphique.

Ainsi, pour afficher le diagramme en bâtons associé à une série statistique brute L , on écrira généralement le code suivant.

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

Exercice 18. Écrire un code Python permettant d'afficher le diagramme en bâtons de l'exercice 4.

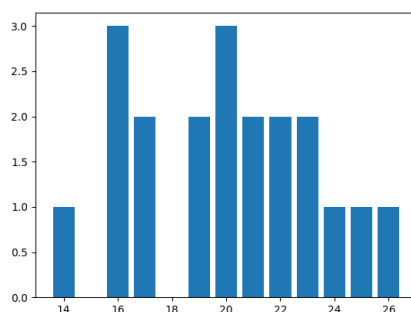
2. Regroupement en classes et histogramme normalisé

Dans certaines études statistiques, le caractère observé est mesuré avec une précision qui peut rendre l'affichage d'un diagramme en bâtons inintéressant.

Prenons, par exemple, ce relevé des tailles (en centimètres) d'une population de rats.

[25, 26, 19, 16, 21, 20, 17, 22, 16, 22, 21, 19, 24, 20, 14, 23, 17, 16, 23, 20]

Son diagramme en bâtons est le suivant.



Sa lisibilité laisse à désirer : voir comptés différemment deux individus de tailles très proches rend le graphique peu digeste.

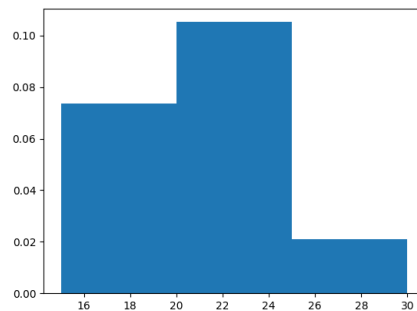
On préférerait avoir un aperçu de la répartition des tailles selon des critères plus grossiers. Par exemple, on pourrait vouloir :

- déclarer des "groupes" de tailles.
- afficher le nombre d'individus de chaque groupe.

Ici, par exemple, on pourrait vouloir considérer trois groupes :

- le groupe des "petits" rats, dont la taille T vérifie $15 \leq T < 20$,
- le groupe des "moyens", dont la taille T vérifie $20 \leq T < 25$,
- le groupe des "plus grands", dont la taille T vérifie $25 \leq T < 30$.

Dans ce cas, on représente les données sous la forme d'un **histogramme normalisé**.



On voit ici plus facilement l'importance de chaque groupe, et on peut adapter nos groupes selon des critères convenables à notre contexte.

Pour lire un histogramme normalisé, il faut comprendre les conventions suivantes :

- Les "groupes" qu'on a considéré sont appelés **des classes**. Un histogramme normalisé est formé d'un rectangle par classe.
- **L'aire de ces rectangles** donne la fréquence de leur classe, c'est-à-dire la fréquence avec laquelle un individu tombe dans la classe en question.

C'est donc **l'aire des rectangles qui est directement pertinente** dans un histogramme normalisé, et non pas les hauteurs (qui indiquent une "densité").

Définition 19. Soit X une variable statistique quantitative.

Un **regroupement en classes** des modalités de X est la donnée d'un entier $q \in \mathbb{N}^*$ 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 $i \in \llbracket 1, q \rrbracket$, on dit que $I_i = [y_i, y_{i+1}[$ est **une classe** de X (pour ce regroupement en classe).

Le vocabulaire relatif aux modalités se transpose aux classes.

Définition 20. Fixons un regroupement en classe

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

de la variable statistique quantitative X étudiée. Pour tout $i \in \llbracket 1, q \rrbracket$:

- On appelle **effectif de la classe** I_i l'entier naturel noté ν_i dans la suite donné par :

$$\nu_i = \text{Card}(\{j \in \llbracket 1, n \rrbracket | X(\omega_j) \in [y_i, y_{i+1}[\}).$$

- On appelle **fréquence de la classe** I_i le nombre réel, noté ϕ_i dans la suite, donné par :

$$\phi_i = \frac{\nu_i}{n}.$$

- On appelle **effectif cumulée de la classe** I_i l'entier, noté ν_i^c dans la suite, donné par :

$$\nu_i^c = \text{Card}(\{j \in \llbracket 1, n \rrbracket | X(\omega_j) \in [y_1, y_{i+1}[\}).$$

- On appelle **fréquence cumulée de la classe** I_i le nombre réel, noté ϕ_i^c dans la suite, donné par :

$$\phi_i^c = \frac{\nu_i^c}{n}.$$

Exercice 21. Pour les données de l'exercice 4, on choisit d'effectuer un regroupement en classe avec les intervalles $[12, 15[$, $[15, 18[$, $[18, 21[$, $[21, 24[$ et $[24, 27[$.

1. Déterminer les effectifs, effectifs cumulés, fréquences et fréquences cumulées de chaque classe.
2. Écrire le code d'une fonction **Regroupe** prenant en entrée une série statistique brute sous la forme d'une liste L et la liste $[y(1), y(2), \dots, y(q+1)]$ des extrémités d'un regroupement en classes, et renvoyant en sortie le couple N, F formé :
 - de la liste $N = [n_1, \dots, n_q]$ des effectifs de ces classes,
 - de la liste $F = [f_1, \dots, f_q]$ des fréquences de ces classes.
3. Tester la fonction **Regroupe** sur l'exemple de l'exercice 4.

Définition 22. Considérons un regroupement en classe $[y_1, y_2[, \dots, [y_q, y_{q+1}[$ de la variable statistique quantitative X étudiée.

L'histogramme normalisé (de la série statistique $(X(\omega_1), \dots, X(\omega_n))$) associé à ce regroupement en classe est l'union, dans le plan des rectangles :

- dont les bases sont les classes $[y_1, y_2[, [y_2, y_3[, \dots$, et $[y_q, y_{q+1}[$,
- et dont les hauteurs sont, respectivement, $\frac{\phi_1}{y_2 - y_1}, \frac{\phi_2}{y_3 - y_2}, \dots, \frac{\phi_q}{y_{q+1} - y_q}$.

L'aire du rectangle dont la base est I_i est donc la fréquence ϕ_i de cette classe (pour tout $i \in \llbracket 1, q \rrbracket$).

Histogramme normalisé en Python

Considérons une série statistique brute et un regroupement en classe $[y_1, y_2[, \dots, [y_q, y_{q+1}[$.

Notons L le vecteur Numpy (ou la liste) contenant la série statistique brute.

Soit C la liste $C = [y(1), \dots, y(q), y(q+1)]$ contenant les extrémités de ces classes.

Alors :

- La commande `plt.hist(L, C, density=True)` engendre l'histogramme normalisé associé à ce regroupement en classe,
- et la commande `plt.show()` demande à Python d'afficher ce graphique.

Exemple 23. L'histogramme normalisé précédent, donnant la taille des rats, a été affiché à l'aide du code suivant.

```
L=[25, 26, 19, 16, 21, 20, 17, 22, 16, 22, 21, 19, 24, 20, 14, 23, 17, 16, 23, 20] #
    Tailles des rats
C=[15, 20, 25, 30] # Extrémités des classes
plt.hist(L, C, density=True)
plt.show()
```

Exercice 24. Écrire un code Python permettant d'afficher l'histogramme normalisé associé à la série statistique de l'exercice 4, selon le regroupement en classes donné par les intervalles $[12, 15[$, $[15, 18[$, $[18, 21[$, $[21, 24[$ et $[24, 27[$.

Exercice 25. Écrire un code Python permettant d'afficher l'histogramme normalisé associé à la série statistique de l'exercice 4, selon un regroupement en classes de 10 classes de la même longueur. On pourra utiliser la fonction `np.linspace`.

3. Diagramme des fréquences cumulées

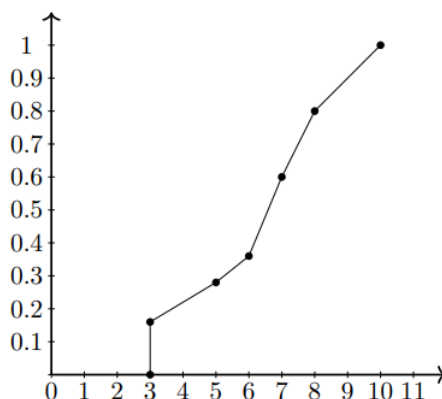
Voici une dernière manière classique de représenter des données.

Définition 26. (Diagramme des fréquences cumulées par modalités)

Le diagramme des fréquences cumulées (par modalités) de la série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$ est la ligne brisée reliant, dans cet ordre les points de coordonnées $(x_1, 0)$, (x_1, f_1^c) , $(x_2, f_2^c), \dots, (x_p, f_p^c)$.

Voici un exemple :

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



On peut recourir à ce type de diagramme en utilisant un regroupement en classes.

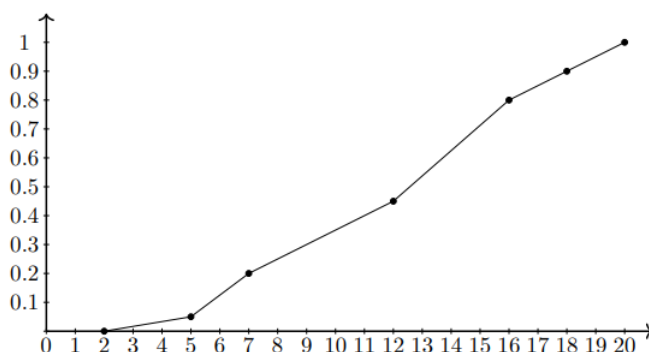
Définition 27. (Diagramme des fréquences cumulées avec regroupement en classes)

Considérons la série statistique dépouillée $((x_1, n_1), \dots, (x_p, n_p))$ et un regroupement en classes $[y_1, y_2[$, $[y_2, y_3[$, ..., et $[y_q, y_{q+1}[$.

Le diagramme des fréquences cumulées associé à ce regroupement en classes est la ligne brisée reliant, dans cet ordre, les points $(y_1, 0)$, (y_2, ν_1^c) , $(y_3, \nu_2^c), \dots, (y_q, \nu_{q-1}^c)$, (y_{q+1}, ν_q^c) .

Et un exemple :

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 28.** 1. Écrire un code Python permettant d'afficher le diagramme des fréquences cumulées (par modalités) de la série statistique de l'exercice 4.
2. Écrire un code Python permettant d'afficher le diagramme des fréquences cumulées de la série statistique de l'exercice 4 selon le regroupement en classes donné par les intervalles $[12, 15[$, $[15, 18[$, $[18, 21[$, $[21, 24[$ et $[24, 27[$.

4. Exercices de synthèse pour cette partie

Exercice 29. 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).

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

1. Écrire un code Python affichant les modalités et effectifs de cette série statistique.
2. Écrire un code Python affichant le diagramme bâton associé à cette série statistique.
3. Écrire un code Python affichant le diagramme des fréquences cumulées associé à cette série statistique.

Exercice 30. 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).

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

1. Écrire un code Python affichant le diagramme des fréquences cumulées associé à cette série statistique.
2. Écrire un code Python affichant le diagramme des fréquences cumulées associé à cette série statistique pour le regroupement en classes défini par la liste $C = [5, 8, 11, 14, 17, 20]$.

III. Indicateurs statistiques

Dans cette partie, on s'intéresse aux indicateurs statistiques classiques : ce sont des quantités qui permettent de décrire une série statistique.

- On dispose d'indicateurs statistiques dits "de position", qui permettent de situer les valeurs observées. Par exemple, la moyenne empirique, la médiane et les modes sont de tels indicateurs.
- On dispose d'indicateurs dits "de dispersion", qui permettent de quantifier la variabilité des valeurs observées. Par exemple, l'étendue et la variance empirique sont de tels indicateurs.

Dans toute cette partie, on considère toujours l'observation d'une variable statistique discrète X sur un échantillon $(\omega_1, \dots, \omega_n)$.

Pour faciliter l'écriture et la lecture :

- On notera $x = (v_1, \dots, v_n) = (X(\omega_1), \dots, X(\omega_n))$ le n -uplet des valeurs observées,
- on notera toujours $x_1, \dots, x_p$ les modalités de X dans l'ordre croissant ($x_1 < \dots < x_p$).
- Pour tout $i \in \llbracket 1, p \rrbracket$, on note toujours n_i, n_i^c, f_i, f_i^c l'effectif, l'effectif cumulé, la fréquence et la fréquence cumulée de la modalité x_i .

Les indicateurs présentés seront illustrés sur la série statistique brute suivante :

$[2.5, 1.5, 3.0, 0.5, 3.0, 2.5, 3.5, 4.0, 4.5, 4.0, 3.5, 2.0, 3.5, 1.5, 0.5, 1.5, 2.5, 4.0, 5.0, 10.5, 12.0, 11.0, 7.5, 10.5, 7.0, 6.0, 5.5, 5.0]$

donnant la liste des températures journalières moyennes observées en février 2025 dans les Yvelines.

1. Indicateurs de position

a) Modes, minimum et maximum

Définition 31. On appelle **mode** d'une série statistique toute modalité d'effectif maximal. Avec nos notations: soit $i \in \llbracket 1, p \rrbracket$. La modalité x_i est un mode de la série statistique x étudiée si pour tout $j \in \llbracket 1, p \rrbracket$, $n_i \geq n_j$.

Exemple 32. Pour les relevés de température, le dépouillement de cette série statistique donne le tableau ci-dessous.

x_i	0,5	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	7	7,5	10,5	11	12
n_i	2	3	1	3	2	3	3	1	2	1	1	1	1	2	1	1

Cette série statistique admet donc 4 modes : les modalités $1,5^\circ C$, $2,5^\circ C$, $3,5^\circ C$, et $4^\circ C$.

Définition 33. Soit $i \in \llbracket 1, n \rrbracket$.

- On dit que la valeur v_i est le **minimum** de la série statistique x si : $\forall j \in \llbracket 1, n \rrbracket, v_i \leq v_j$.
- On dit que la valeur v_i est le **maximum** de la série statistique x si : $\forall j \in \llbracket 1, n \rrbracket, v_i \geq v_j$.

Ainsi, le minimum de la série statistique x est la modalité x_1 , et son maximum est la modalité x_p .

Exemple 34. Le minimum de la série statistique des relevés de température de février 2025 est $0,5^\circ C$, son maximum est $12^\circ C$.

b) Moyenne empirique et médiane

Définition 35. La **moyenne** (empirique) de la série statistique $x = (v_1, \dots, v_n)$ est le réel noté $\bar{x}$ défini par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n v_i.$$

Exemple 36. Pour la série statistique x donnée par les relevés de température, on trouve $\bar{x} \simeq 4,57$.

Proposition 37. On a les égalités :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^p n_i x_i = \sum_{i=1}^p f_i x_i.$$

Démonstration. En exercice. $\square$

Remarque. Pour décrire l'égalité $\bar{x} = \sum_{i=1}^p f_i x_i$, on dit que la moyenne de la série statistique x est la moyenne des modalités pondérées par leur fréquence.

Définition 38. La **médiane** de la série statistique x est le réel noté M_e donné par

$$M_e = x_{i_0}$$

où $i_0 = \min\{i \in \llbracket 1, p \rrbracket, f_i^c \geq \frac{1}{2}\}$.

Autrement dit, la médiane d'une série statistique x est la plus petite modalité dont la fréquence cumulée dépasse $\frac{1}{2}$.

Exemple 39. Le calcul des fréquences cumulées pour le relevé des températures donne le tableau suivant.

x_i	0,5	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	7	7,5	10,5	11	12
$f_i^c \simeq$	0,071	0,178	0,214	0,321	0,392	0,5	0,607	0,642	0,714	0,75	0,785	0,821	0,857	0,928	0,964	1

La médiane est donc la modalité $M_e = 3,5^\circ C$.

Ici, cette modalité a une fréquence cumulée d'exactement $1/2$, ce qui est un hasard.

Proposition 40. La médiane M_e d'une série statistique x partage les valeurs observées au sens suivant :

- Plus de la moitié des valeurs observées sont supérieures ou égales à M_e :

$$\text{Card}\{j \in \llbracket 1, n \rrbracket \mid v_j \geq M_e\} \geq n/2.$$

- plus de la moitié des valeurs observées sont inférieures ou égales à M_e :

$$\text{Card}\{j \in \llbracket 1, n \rrbracket \mid v_j \leq M_e\} \geq n/2.$$

Démonstration. En exercice. $\square$

Remarque. Étant donné une série statistique brute donnée sous la forme d'une liste Python L , pour déterminer la médiane on peut :

- Effectuer un tri croissant de L ,
- prendre la valeur du milieu de la liste triée si celle-ci est de longueur impaire, et la valeur juste avant le milieu si celle-ci est de longueur paire.

Ainsi, soit $(v'_1, \dots, v'_n)$ est la série statistique obtenue en triant dans l'ordre croissant la série statistique $x = (v_1, \dots, v_n)$. On dit que $(v'_1, \dots, v'_n)$ est la série statistique **ordonnée** associée à x .

La médiane de x est alors la valeur v'_k où :

$$k = \begin{cases} \frac{n+1}{2} & \text{si } n \text{ impair} \\ \frac{n}{2} & \text{sinon} \end{cases}.$$

On pourra remarquer qu'alors, $k = \lfloor \frac{n+1}{2} \rfloor$ (faire une disjonction des cas).

c) Quartiles, boîte à moustache

Définition 41.

- Le **premier quartile** de la série statistique $x = (v_1, \dots, v_n)$, noté q_1 dans la suite, est la plus petite modalité pour laquelle la fréquence cumulée dépasse $\frac{1}{4}$.
- Le **troisième quartile** de la série statistique $x = (v_1, \dots, v_n)$, noté q_3 dans la suite, est la plus petite modalité pour laquelle la fréquence cumulée dépasse $\frac{3}{4}$.

Exemple 42. Reprenons le tableau des fréquences cumulées pour le relevé des températures.

x_i	0,5	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	7	7,5	10,5	11	12
$f_i^c \simeq$	0,071	0,178	0,214	0,321	0,392	0,5	0,607	0,642	0,714	0,75	0,785	0,821	0,857	0,928	0,964	1

Le premier quartile est la modalité $q_1 = 2,5^\circ C$.

Le troisième quartile est la modalité $q_3 = 5,5^\circ C$.

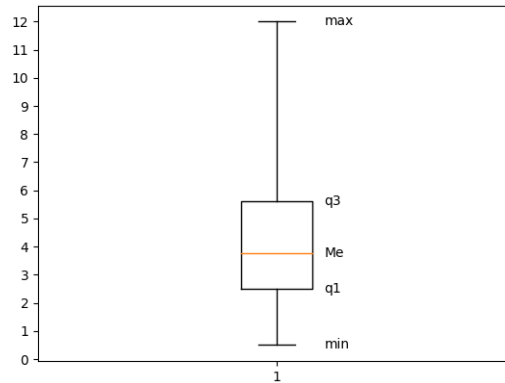
d) Boîte à moustache

Une **boîte à moustache** est un diagramme résumant les indicateurs de position suivants : minimum, premier quartile, médiane, troisième quartile et maximum.

Pour le relevé des températures, nous avons :

- Minimum : 0, 5.
- Premier quartile : $q_1 = 2,5$, médiane : $M_e = 3,5$, Troisième quartile : 5, 5.
- Maximum : 12.

La boîte à moustache de cette série statistique se présente sous la forme suivante :



Boîte à moustache en Python

Étant donné une série statistique brute donnée en Python par un tableau de type `array` ou une liste `L`, les lignes suivantes (utilisant `matplotlib.pyplot`):

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

génèrent et affichent la boîte à moustache de cette série statistique.

e) Quantile d'ordre α

Définition 43. Soit $\alpha \in]0, 1[$. Le **quantile d'ordre α** de la série $x = (v_1, \dots, v_n)$, noté Q_α est la plus petite modalité dont la fréquence cumulée dépasse α .

Ainsi, $Q_\alpha = x_k$ où :

$$k = \min\{i \in \llbracket 1, p \rrbracket \mid f_i^c \geq \alpha\}.$$

Exemple 44. • Le premier quartile est le quantile d'ordre $\frac{1}{4}$. La médiane est le quantile d'ordre $\frac{1}{2}$.

- Soit $k \in \llbracket 1, 9 \rrbracket$. On appelle **k -ième décile** de x , noté D_k , le quantile d'ordre $\frac{k}{10}$ de x .
- L'écart inter-décile est la distance $D_9 - D_1$ entre le premier et le dernier décile.
- Soit $k \in \llbracket 1, 99 \rrbracket$. On appelle **k -ième centile** de x , noté C_k , le quantile d'ordre $\frac{k}{100}$ de x . L'écart inter-centile est $C_{99} - C_1$.

Par exemple, pour les relevés de température...

x_i	0,5	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	7	7,5	10,5	11	1
$f_i^c \simeq$	0,071	0,178	0,214	0,321	0,392	0,5	0,607	0,642	0,714	0,75	0,785	0,821	0,857	0,928	0,964	1

- Le premier décile est $D_1 = 1,5^\circ C$. Le 7e décile est $D_7 = 5^\circ C$. Le 9e décile est $D_9 = 10,5^\circ C$. L'écart inter-décile est de $D_9 - D_1 = 9^\circ C$.
- Le 76e centile est la première modalité dont l'effectif cumulé dépasse 0,76, donc c'est ici $C_{76} = 6^\circ C$.

2. Indicateurs de dispersion

a) Étendue, écart inter-quartile

Définition 45. L'**étendue** de la série statistique x est la différence $x_p - x_1$ entre la plus grande modalité et sa plus petite. c'est donc la longueur de l'intervalle dans lequel x prend ses valeurs.

Définition 46. L'**écart inter-quartile** de la série x est la différence $q_3 - q_1$ entre le troisième et le premier quartile.

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

Remarque. L'intervalle inter-quartile contient plus de la moitié des valeurs :

Proposition 47. $\text{Card}\{j \in \llbracket 1, n \rrbracket | v_j \in [q_1, q_3] \geq \frac{n}{2}.$

Démonstration. En exercice. $\square$

Exemple 48. L'étendue et l'écart inter-quartile sont visibles sur une boîte à moustache.

Nous avons, pour les relevés de température :

- Minimum : 0, 5.
- Premier quartile : $q_1 = 2, 5$, médiane : $M_e = 3, 5$, Troisième quartile : 5, 5.
- Maximum : 12.

L'étendue est donc de 11, 5, l'écart inter-quartile de 3.

b) Variance et écart-type

Définition 49. • On appelle **variance** (empirique) de la série statistique x , et on note s_x^2 , le réel (positif) donné par :

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (v_i - \bar{x})^2.$$

- **L'écart-type** (empirique) de x est le réel noté σ_x donné par :

$$\sigma_x = \sqrt{s_x^2}.$$

Remarque. • La variance mesure donc la moyenne des *écart quadratiques* entre les valeurs de la série statistiques et sa moyenne. On parle d'écart quadratique, car on met la distance $|v_i - \bar{x}|$ de v_i à $\bar{x}$ au carré.

- La variance est positive en tant que somme des réels positifs $\frac{(v_i - \bar{x})^2}{n}$. Donc σ_x est bien défini.

Proposition 50. Pour toute série statistique x , on a l'égalité :

$$s_x^2 = \frac{1}{n} \sum_{i=1}^p n_i (x_i - \bar{x})^2 = \sum_{i=1}^p f_i (x_i - \bar{x})^2.$$

Démonstration. En exercice, corrigé à noter. $\square$

On a la très importante formule de Koenig-Huygens donnant d'autres manières de calculer la variance empirique.

Proposition 51. Notons x^2 la série statistique $(v_1^2, \dots, v_n^2)$ obtenue en mettant au carré les valeurs de x . $\overline{x^2}$ désigne donc la moyenne empirique de cette série statistique. On a alors les résultats suivants :

- $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n v_i^2 = \frac{1}{n} \sum_{i=1}^p n_i x_i^2 = \sum_{i=1}^p f_i x_i^2.$
- $s_x^2 = \overline{x^2} - (\bar{x})^2.$

Exemple 52. Pour la série statistique x suivante :

4	5	4	6	3	4	7	4	3	6	2
---	---	---	---	---	---	---	---	---	---	---

Le dépouillement de x est résumé par le tableau suivant :

x_i	2	3	4	5	6	7
n_i	1	2	4	1	2	1

Son effectif total étant 11, on a :

$$\overline{x^2} = \frac{1 \cdot 2^2 + 2 \cdot 2^2 + 4 \cdot 4^2 + 1 \cdot 5^2 + 2 \cdot 6^2 + 1 \cdot 7^2}{11} \simeq 21,1$$

On calcule de même la moyenne empirique de x :

$$\bar{x} \simeq 4,36$$

Alors, on a :

$$s_x^2 = \overline{x^2} - (\bar{x})^2 \simeq 21,1 - (4,36)^2 \simeq 2,09.$$

Puis :

$$\sigma_x = \sqrt{s_x^2} \simeq 1,44.$$

c) Fonction de répartition empirique

Définition 53. Soit x une série statistique brute.

La **fonction de répartition empirique** de x est la fonction notée F_x donnée par :

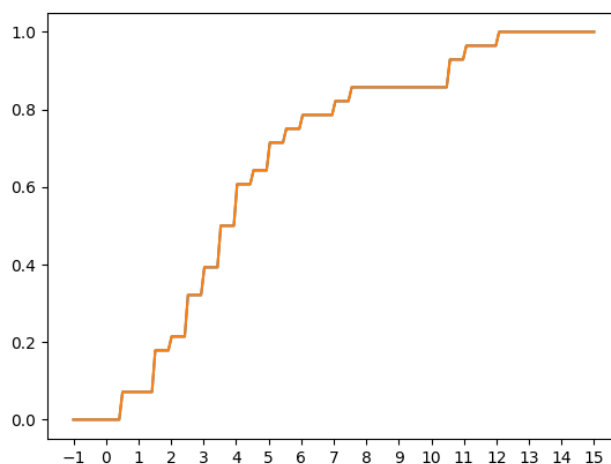
$$F_x : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ t & \longmapsto \frac{\text{Card}\{j \in \llbracket 1, n \rrbracket \mid v_j \leq t\}}{n} \end{cases}.$$

Autrement dit, pour tout réel t , $F_x(t)$ est la fréquence avec laquelle les valeurs de x sont inférieures à t .

Remarque. Ainsi, pour tout réel t :

- Si $t < x_1$, alors $F_x(t) = 0$.
- Si on dispose $i \in \llbracket 1, p-1 \rrbracket$ tel que $t \in [x_i, x_{i+1}[$, alors $F_x(t) = f_i^c$ (effectif cumulé).
- Si $t \geq x_p$, alors $F_x(t) = 1$.

Exemple 54. Pour les relevés de température x , sa fonction de répartition F_x a le graphe suivant. On remarque que ses discontinuités sont les modalités de x .



IV. Moyenne et variance : aspects théoriques

1. Propriétés de la moyenne

Ces énoncés sont importants, et leurs démonstrations en exercices sont élémentaires mais nous serons amenés à les reproduire dans un cadre théorique plus construit.

Ici, "série statistique" désigne par défaut une série statistique brute.

Proposition 55. (Stabilité de la moyenne)

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

- Supposons donnés des réels a et b tels que : $\forall i \in \llbracket 1, n \rrbracket, a \leq v_i \leq b$. Alors :

$$a \leq \bar{x} \leq b.$$

- En particulier :

$$x_1 \leq \bar{x} \leq x_p.$$

Démonstration. En exercice (sommation d'inégalités). $\square$

Définition 56. Soit $x = (v_1, \dots, v_n)$ une série statistique. On dit qu'une série statistique y est une **transformée affine** de x s'il existe des réels a et b tels que :

$$y = (av_1 + b, av_2 + b, \dots, av_n + b).$$

Pour de tels réels a et b , on note alors $y = ax + b$.

Proposition 57. (linéarité de la moyenne)

Soient x une série statistique, et y une transformée affine de x .

Soient a et b deux réels tels que $y = ax + b$.

Alors, $\bar{y} = a\bar{x} + b$.

Démonstration. En exercice (linéarité de la somme). $\square$

Définition 58. Soit x une série statistique. On dit que x est **centrée** si $\bar{x} = 0$.

Proposition 59. Soit x une série statistique. Alors, la transformée affine $x - \bar{x}$ est centrée.

Démonstration. En exercice (utiliser la proposition précédente). $\square$

2. Propriétés de la variance

Voici une propriété fondamentale de la variance.

Proposition 60. (Séries statistiques de variance nulle)

Soit $x = (v_1, \dots, v_n)$ une série statistique. Il est équivalent de dire :

- $s_x^2 = 0$, et
- x est constante : $v_1 = v_2 = \dots = v_n$.

Démonstration. En exercice (une somme de réels positifs est nulle si et seulement si tous ses termes sont nuls). $\square$

Proposition 61. Soient x une série statistique, y une transformée affine de x et a, b des réels tels que $y = ax + b$.

Alors :

- $s_y^2 = a^2 s_x^2$.
- $\sigma_y = |a| \sigma_x$.

Démonstration. En exercice (linéarité de la somme et de la moyenne empirique). $\square$

Une conséquence simple :

Définition 62. On dit qu'une série statistique x est **centrée réduite** si $\bar{x} = 0$ et $s_x^2 = 1$.

Proposition 63. Soit x une série statistique non constante. Alors, $\frac{x - \bar{x}}{\sigma_x}$ est une série statistique centrée réduite.

Remarque. Dans cette proposition, $\frac{x - \bar{x}}{\sigma_x}$ est la transformée affine $\frac{1}{\sigma_x}x - \frac{\bar{x}}{\sigma_x}$ de x .

Démonstration. En exercice. $\square$

V. Python : exercices et commandes au programme

1. Retour sur les parties précédentes

Exercice 64. Écrire le code d'une fonction Python prenant en entrée une série statistique brute donnée sous la forme d'un vecteur numpy de type `array`, et renvoyant en sortie ce qui est indiqué.

On pourra utiliser la commande `np.sum(A)` (au programme) renvoyant la somme des valeurs d'un vecteur numpy `A`, mais pas les fonctions présentées dans le paragraphe suivant.

1. La liste des modes de la série donnée en entrée.
2. Le minimum de la série donnée en entrée.
3. Le maximum de la série donnée en entrée.
4. La moyenne de la série donnée en entrée.
5. La médiane de la série donnée en entrée.
6. Le couple (q_1, q_3) formé du premier et du troisième quartile de la série donnée en entrée.
7. L'étendue de la série donnée en entrée.
8. L'écart inter-quartile de la série donnée en entrée.
9. La variance de la série donnée en entrée.
10. L'écart type de la série donnée en entrée.

Vérifier vos fonctions à l'aide de la série statistique utilisée dans le cours, donnant les relevés de température de février 2025 dans les Yvelines, et les exemples donnés.

2. Indicateurs statistiques avec Numpy

Commandes au programme

Soit `A` un vecteur numpy (de type `array`) contenant des valeurs numériques. Alors :

- `np.min(A)` renvoie le minimum de `A`.
- `np.max(A)` renvoie le maximum de `A`.
- `np.mean(A)` renvoie la moyenne des valeurs de `A`.
- `np.median(A)` renvoie la médiane des valeurs de `A` selon la définition alternative ci-dessous.
- `np.var(A)` renvoie la variance des valeurs de `A`.
- `np.std(A)` renvoie l'écart-type des valeurs de `A`.

Remarque. Définition alternative de la médiane (aucune ne fait consensus) : Définition alternative de la médiane utilisée par Python :

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

- Si n est impair, on dispose de k entier tel que $n = 2k + 1$ et la médiane M_e est toujours la valeur centrale $M_e = v'_{k+1}$.
- Si n est pair, on dispose de k entier tel que $n = 2k$ et la médiane est alors la moyenne $M_e = \frac{v'_k + v'_{k+1}}{2}$ des deux valeurs centrales de la série ordonnée.

3. Exercices

Exercice 65. Écrire le code d'une fonction Python d'entête `def F(x, t):` prenant en entrée un vecteur numpy x représentant une série statistique brute x ainsi qu'un réel t , et renvoyant en sortie la valeur de $F_x(t)$ où F_x est la fonction de répartition empirique de x .

Exercice 66. 1. Déterminer la valeur de tous les indicateurs statistiques de l'exercice 64, pour les séries statistiques des exercices 29 et 30. Tracer les boîtes à moustache de ces séries statistiques.

2. Tracer le graphe de la fonction de répartition empirique de la série statistique de l'exercice 29.

Exercice 67. 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. Cette série est donnée en Python par :

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

1. Déterminer les indicateurs statistiques mentionnés dans l'exercice 64 de cette série statistique.
2. Représenter ces données sous forme d'un diagramme des fréquences cumulées, puis sous la forme d'une boîte à moustache.
3. Déterminer l'intervalle inter-décile et inter-centile de cette série statistique.

Exercice 68. On considère la série statistique brute contenant un échantillon de 100 salaires mensuels nets des habitants de la rue des orangers. Cette série est donnée en Python par :

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

On regroupe ces données en 5 classes selon les valeurs 1000, 1500, 2000, 2500, 3000 et 3500.

1. Afficher le diagramme des fréquences cumulées de cette série statistique pour ce regroupement en classe.
2. Afficher l'histogramme normalisé de cette série statistique pour ce regroupement en classe.

Exercice 69. Soit $x = (v_1, \dots, v_n)$ une série statistique.

On pose, pour tout réel t , $f(t) = \sum_{i=1}^p f_i(x_i - t)^2$.

Montrer que la fonction f ainsi définie admet un unique extremum local sur $\mathbb{R}$, et que celui-ci est un extremum global. Déterminer sa valeur m ainsi que celle de son unique antécédent par f .

Exercice 70. Soient $x = (v_1, \dots, v_n)$ et $y = (w_1, \dots, w_m)$ deux séries statistiques brutes.

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

Montrer que $\bar{z} = \frac{n\bar{x} + m\bar{y}}{n + m}$.