TP 7 : Statistiques univariées linéaires

Ce chapitre est à mettre en parallèle avec le TP `Python` sur les statistiques.

I) Vocabulaire

Les premières études statistiques étaient des recensements démographiques : on en a conservé le vocabulaire.

Définition 1

L'ensemble des éléments dont on étudie les données s'appelle **population**. Classiquement, on notera cet ensemble Ω. Un élément de Ω est appelé individu et se note ω.

Un échantillon est une liste finie d'individus sur lequel on effectue des observations. On appelle taille de l'échantillon, le nombre d'individus, on le note N.

Définition 2

On appelle **caractère** (ou encore **variable statistique**) l'aspect que l'on observe sur les individus. Lorsque les différentes valeurs d'un caractère sont des nombres, le caractère est *quantitatif*. Dans le cas contraire, le caractère est *qualitatif*.
Si on note X la variable statistique, on note $X(\Omega)$ l'ensemble des valeurs prises par cette variable.
On dit que la variable X est discrète si $X(\Omega)$ est finie ou dénombrable. Dans le cas contraire, on dit qu'elle est continue.

Remarque : Dans la suite, on se concentrera sur les caractères quantitatives.

Définition 3

Les valeurs prises par une variable statistique X discrète s'appellent modalités de X. Dans la suite, nous les noterons $x_1,\dots,x_p$ avec $x_1 < \dots < x_p$.

Si X est continue, on regroupe les valeurs dans des intervalles que l'on appelle classes de X. Nous les noterons $[x_1,x_2[, [x_2,x_3[,\dots,[x_p,x_{p+1}[$ lorsqu'il y en a p.

Remarque : Il est possible de regrouper les valeurs de X dans des intervalles même dans le cas discret.

Définition 4 (Effectifs)

On appelle effectif de la modalité x_i ou de la classe $[x_i,x_{i+1}[$, noté n_i, le nombre d'individus ω de l'échantillon tels que $X(\omega) = x_i$ ou $X(\omega) \in [x_i,x_{i+1}[$.

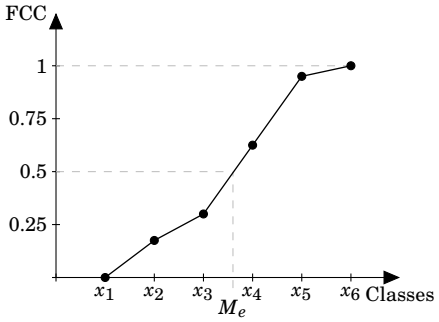
On appelle effectif cumulé croissant en x_i (ou en $[x_i,x_{i+1}[$) la somme des effectifs des modalités (ou des classes) qui lui sont inférieures ou égales.

Définition 5 (Fréquences)

On appelle fréquence d'une valeur est le quotient $f_i = \dfrac{n_i}{N}$.

On appelle fréquence cumulée croissante en x_i (ou en $[x_i,x_{i+1}[$) la somme des fréquences des modalités (ou des classes) qui lui sont inférieures ou égales.

Remarque : Il peut arriver que l'on représente les fréquences cumulées croissantes sur un graphique :



Ce graphique peut notamment être pratique pour repérer la médiane et les quantiles (ces notions seront présentées dans la suite).

Définition 6 (Série statistique)

On appelle série statistique simple d'un échantillon la donnée des modalités accompagnés des effectifs. On note une telle série statistique $(x_i, n_i)_{1\leq i\leq p}$.
Lorsque les valeurs de la variable sont groupés par classes, on parle de série statistique groupée et on note $([x_i, x_{i+1}[, n_i)_{1\leq i\leq p}$.

Dans la suite de ce chapitre, on n'étudiera que des séries statistiques quantitatives et discrètes.

II) Paramètres de position

1) La moyenne

On considère la série statistique $(x_i, n_i)_{1\leq i\leq p}$.

Définition 7 (Cas discret)

On appelle moyenne de la série statistique $(x_i, n_i)_{1\leq i\leq p}$ le réel noté $\overline{x}$ de la somme de toutes les valeurs de cette série par l'effectif total :

$$\overline{x} = \frac{n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p}{N} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i$$

Remarque : On rappelle que la fréquence de la valeur x_i est $f_i = \dfrac{n_i}{N}$. On peut alors calculer la moyenne à partir des fréquences grâce à la formule $\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_p x_p$, ce qui rappelle fortement la formule de l'espérance d'une variable aléatoire.

Exemple : *Considérons la série statistique suivante :*

Nombre d'interventions x_i	*3*	*5*	*6*	*7*	*8*	*9*
Nombre de jours n_i	*2*	*5*	*8*	*6*	*3*	*1*
Fréquence f_i	*0,08*	*0,2*	*0,32*	*0,24*	*0,12*	*0,04*

Le nombre moyen d'interventions par jour est :

$$\overline{x} = \frac{2\times 3 + 5\times 5 + 8\times 6 + 6\times 7 + 3\times 8 + 1\times 9}{25} = 6,16$$

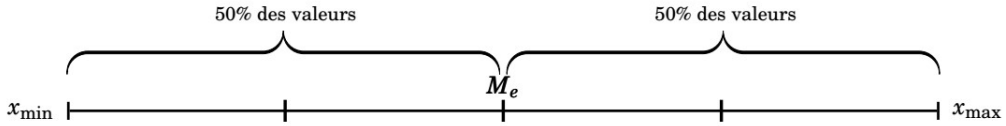
ou en utilisant les fréquences :

$$\overline{x} = 0,08\times 3 + 0,2\times 5 + 0,32\times 6 + 0,24\times 7 + 0,12\times 8 + 0,04\times 9 = 6,16$$

2) La médiane

Définition 8

On appelle médiane d'une série statistique, une valeur qui partage la série statistique en deux de sorte qu'il y ait autant d'observations ayant une valeur supérieure à la médiane que d'observations ayant une valeur inférieure à la médiane.
Généralement, la médian est notée M_e.



Point méthodologique 1

La médiane d'une série statistique de N valeurs rangées par ordre **croissant** est le nombre M_e défini par :

- si l'effectif N est impair, la médiane M_e est la valeur centrale du caractère c'est à dire la valeur de rang $\dfrac{N+1}{2}$ de la série ordonnée.
- si l'effectif N est pair, la médiane M_e est la demi-somme des deux valeurs centrales du caractère c'est à dire la moyenne des valeurs de rangs $\dfrac{N}{2}$ et $\dfrac{N}{2}+1$ de la série ordonnée.

Exemple : *Dans un service de maintenance, on a répertorié le nombre d'interventions par jour sur un mois. On a obtenu la distribution suivante :*

Nombre d'interventions x_i	*3*	*5*	*6*	*7*	*8*	*9*
Nombre de jours n_i	*2*	*5*	*8*	*6*	*3*	*1*

L'effectif total $N = 25$ donc la médiane est la valeur du caractère de rang 13 soit $M_e = 6$.

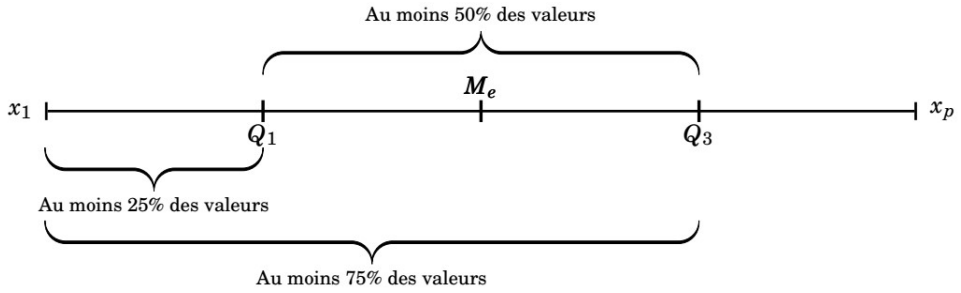
3) Les quantiles

3).1 Les quartiles

Définition 9

— Le premier quartile noté Q_1 est la plus petite valeur de la série statistique telle qu'au moins 25 % des valeurs de la série sont inférieures ou égales à Q_1.
— Le troisième quartile noté Q_3 est la plus petite valeur de la série statistique telle qu'au moins 75 % des valeurs de la série sont inférieures ou égales à Q_3.

Les quartiles au nombre de trois Q_1, Q_2 et Q_3 partagent l'ensemble étudié de N éléments préalablement classés par valeurs **croissantes**, en quatre sous ensembles.



Remarque : L'intervalle interquartile $[Q_1;Q_3]$ contient au moins 50% des valeurs de la série.

Point méthodologique 2

- Pour déterminer le premier quartile Q_1 d'une série statistique, on commence par calculer $\frac{N}{4}$ que l'on arrondie à l'entier supérieur. On note la valeur obtenue N_1. Le premier quartile est la N_1ième valeur de la série.
- Pour déterminer le troisième quartile Q_3 d'une série statistique, on calcule $\frac{3N}{4}$ que l'on arrondie à l'entier supérieur. On note la valeur obtenue N_3. Le premier quartile est la N_3ième valeur de la série.

Exemple : *Dans la série précédente, l'effectif total* $N = 25$.

- $25 \times \frac{1}{4} = 6,25$ *donc le premier quartile est la valeur du caractère de rang 7 (plus petit entier supérieur à* $N/4$*), soit* $Q_1 = 5$.
- $25 \times \frac{3}{4} = 18,75$ *donc le troisième quartile est la valeur du caractère de rang 19 (plus petit entier supérieur à* $3N/4$*), soit* $Q_3 = 7$.

3).2 Les déciles

Les déciles au nombre de neuf D_1, D_2, …, D_9 partagent l'ensemble étudié de N éléments préalablement classés par valeurs **croissantes**, en dix sous ensembles.

Propriété 1

- Le premier décile noté D_1 est la plus petite valeur de la série statistique telle qu'au moins 10% des valeurs de la série sont inférieures ou égales à D_1.
- Le neuvième décile noté D_9 est la plus petite valeur de la série statistique telle qu'au moins 90% des valeurs de la série sont inférieures ou égales à D_9.

La méthode pour trouver les déciles est la même que pour les quartiles en changeant les valeurs des pourcentages recherchés.

III) Caractéristiques de dispersion

La moyenne et les quantiles permettent d'obtenir des informations sur la position d'une série statistique mais elle ne permet aucune interprétation sur les écarts des valeurs de la série statistique. Deux classes peuvent avoir la même moyenne mais la première peut avoir un niveau très homogène lorsque la seconde est hétérogène. Les notions présentées dans cette partie servent à trouver ce genre d'information.

Propriété 2

- **L'étendue** est la différence entre la plus grande et la plus petite valeur d'une série statistique.
- **L'écart interquartile** est égal à la différence entre le troisième et le premier quartiles.
- **L'écart interdécile** est égal à la différence entre le neuvième et le premier déciles.

Variance et écart-type

Définition 10

Soit $(x_i;n_i)$, $1 \leqslant i \leqslant p$, une série statistique de moyenne $\overline{x}$ et d'effectif total N.

- La variance de cette série est le nombre s_x^2 défini par :

$$s_x^2 = \frac{n_1 \times (x_1 - \overline{x})^2 + n_2 \times (x_2 - \overline{x})^2 + \cdots + n_p \times \left(x_p - \overline{x}\right)^2}{N} = \frac{1}{N} \sum_{i=1}^{p} n_i (x_i - \overline{x})^2$$

- L'écart-type, noté s_x, de cette série est égal à la racine carrée de la variance :

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

Remarque : Les valeurs $(x_i - \overline{x})$ sont les « écarts à la moyenne » ; les « carrés des écarts à la moyenne » sont donc $(x_i - \overline{x})^2$.

En faisant la moyenne des carrés des écarts à la moyenne, on trouve la variance.

La variance est donc la moyenne des carrés des écarts à la moyenne $\overline{x}$.

Propriété 3 (Variance empirique)

Pour un n-uplet $(x_1, x_2, \ldots, x_n)$, la variance est :

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

Exemple : *Dans la série précédente de moyenne $\overline{x} = 6,16$ la variance est :*

$$s_x^2 = \frac{2\times(3-6,16)^2 + 5\times(5-6,16)^2 + 8\times(6-6,16)^2 + 6\times(7-6,16)^2 + 3\times(8-6,16)^2 + 1\times(9-6,16)^2}{25} = 1,9744$$

L'écart-type de cette série est donc $s_x = \sqrt{1,9744} \approx 1,4$.

Propriété 4 (Formule de Kœnig-Huygens)

Soit $(x_i ; n_i)$, $1 \leqslant i \leqslant p$, une série statistique de moyenne $\overline{x}$ et d'effectif total N.
La variance de cette série est le nombre V défini par :

$$s_x^2 = \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - \overline{x}^2 = \frac{1}{N}\sum_{i=1}^{p} n_i {x_i}^2 - \overline{x}^2$$

IV) Représentation graphique d'une série statistique

Dans cette partie, nous allons présenter les différentes façons de représenter graphiquement une série statistique. Nous verrons comment le faire avec `Python` et le module `matplotlib.pyplot`.

Les diagrammes en bâtons

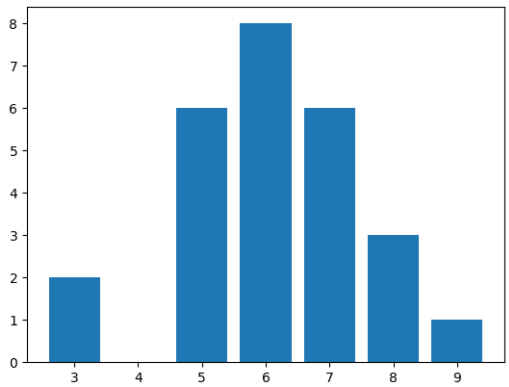
On considère $(x_i, n_i)_{1\leq i\leq p}$ une série statistique. On la représente sur un diagramme en bâton en plaçant les x_i sur un axe horizontal et en représentant à la vertical un bâton de hauteur égale à l'effectif de x_i.

Remarque : On peut faire un diagramme en bâton en représentant les fréquences en ordonnées plutôt que les effectifs. Cela permet notamment de se rapprocher d'une distribution de probabilités.

Exemple : *Reprenant l'exemple du service de maintenance :*

Nombre d'interventions x_i	*3*	*5*	*6*	*7*	*8*	*9*
Nombre de jours n_i	*2*	*5*	*8*	*6*	*3*	*1*

En voici la représentation en bâton :



Remarque : Dans le cas des séries groupées par classes, on préfère utiliser les histogrammes aux diagrammes en bâtons (voir le TP 6 : Le module `matplotlib.pyplot`).

En `Python`, pour faire un diagramme en bâton, on utilise la fonction `plt.bar` (après avoir importer le module `matplotlib.pyplot` comme `plt`). Pour cette représentation, on a besoin d'une liste `x` contenant les modalités de la série statistique et d'une liste `h` contenant les effectifs (ou les fréquences) associées aux modalités de même rang. Par exemple, pour l'exemple précédent, on obtient le diagramme en tapant les lignes suivantes :

Python

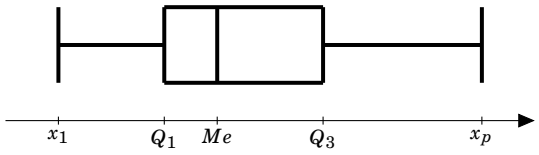
```
import matplolib.pyplot as plt
x=[3,5,6,7,8,9]
y=[2,6,8,6,3,1]
plt.bar(x,y)
plt.show()
```

Remarque : la fonction `bar` possède un grand nombre d'options pour rendre les schémas plus lisibles.

Diagramme en boîte

La représentation graphique de la dispersion d'une série statistique se fait à l'aide de diagramme en boîte appelée aussi « boîte à moustaches ».

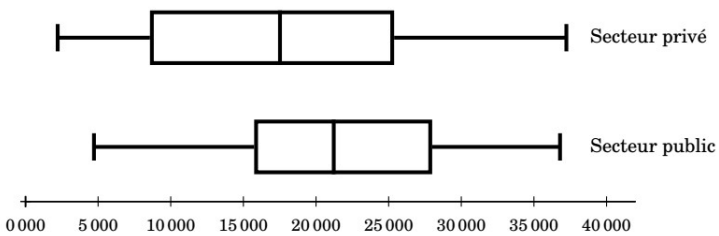
Pour une catégorie donnée, on construit, en face d'un axe permettant de repérer les quantiles de la variable étudiée, un rectangle dont la longueur est égale à l'écart interquartile $Q_3 - Q_1$, la médiane est représentée par un trait. On ajoute alors des segments aux extrémités menant jusqu'aux valeurs extrêmes.



Exemple :
Le tableau suivant donne la distribution du revenu salarial par secteur d'activité en France en 2014.

	Valeur min	Q1	Médiane	Q3	Valeur max
Secteur privé	2218	8570	17520	25377	37234
Secteur public	4716	15744	21221	27996	36797

Source : INSEE

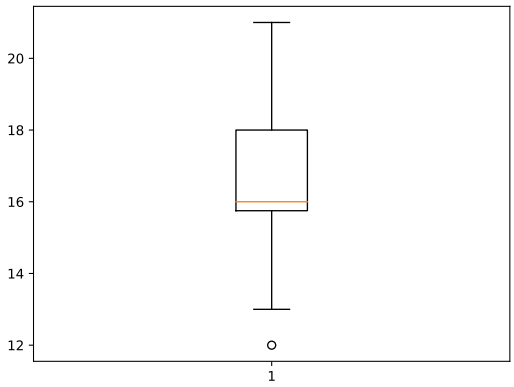


Remarque : Il est possible de représenter des diagrammes en boîtes en utilisant `Python` en utilisant le module `matplotlib.pyplot` et de la fonction `boxplot`. Voici un exemple :

Python

```
import matplolib.pyplot as plt
import random as rd
x=rd.randint(12,22,20)
plt.boxplot(x)
plt.show()
```

Voici le résultat :



V) Transformation affine

Propriété 5

Soit une série statistique dont la moyenne est $\bar{x}$, de variance s_x^2. Considérons la série statistique obtenue en transformant chaque x_i en $ax_i + b$, où a et b sont deux réels, en conservant les mêmes effectifs.

- La moyenne de la nouvelle série statistique est $a\bar{x} + b$.
- La variance de la nouvelle série statistique est $a^2 s_x^2$.

VI) Statistiques et `Python`

Dans cette partie, on présente les commandes qui permettent de calculer les paramètres de positions et de dispersions présentés précédemment.Pour cela, on importe le module numpy :

```
import numpy as np
```

On considère une liste `x` contenant toutes les valeurs de la série statistique avec répétition.

- `np.min(x)` renvoie le plus petit élément de `x` ;
- `np.max(x)` renvoie le plus grand élément de `x` ;
- `np.mean(x)` renvoie la moyenne des éléments de `x` ;
- `np.median(x)` renvoie la médiane des éléments de `x` ;
- `np.var(x)` renvoie la variance des éléments de `x` ;
- `np.std(x)` renvoie l'écart-type des éléments de `x`.

Remarque : En général la liste `x` sera extraite de données statistiques publiques obtenues dans un fichier. Pour faire cela, on a besoin de la bibliothèque `pandas`. Le prochain TP a pour but d'apprendre à travailler avec ce module.

TP 7 : Statistiques univariées linéaires

Vrai ou Faux

Questions	Réponses
1. La médiane d'une série statistique est toujours une valeur de la série	☐ Vrai ☐ Faux
2. Lors du dernier DS de math, la moitié de la classe a eu en dessous de 5 sur 20. La moyenne de la classe est forcément en dessous de 5.	☐ Vrai ☐ Faux
3. L'écart-type est toujours inférieur à la variance.	☐ Vrai ☐ Faux
4. La variance est toujours positive	☐ Vrai ☐ Faux
5. Si on augmente la plus grande valeur d'une série statistique, la médiane et les quartiles ne changent pas.	☐ Vrai ☐ Faux

Exercice 1

Dans une entreprise, il y a 28 cadres et 92 ouvriers. Le salaire moyen des cadres est de 3450 euros et celui des ouvriers est de 1320 euros.

1) Calculer le salaire moyen de l'ensemble des salariés de cette entreprise.

2) a) Quel est le pourcentage d'augmentation du salaire moyen si on verse une prime de 35 euros à chaque salarié ?

b) On augmente le salaire de chaque cadre de 2 % et celui de chaque ouvrier de 4 %.
Le salaire moyen dans l'entreprise a-t-il augmenté de 3% ?

Exercice 2

Un concours est organisé dans deux centres d'examens. Dans le premier centre, les garçons ont obtenu 13 de moyenne et les filles 12 de moyenne. Dans le second centre, les garçons ont obtenu 9 de moyenne et les filles 8 de moyenne. Il y avait 58 garçons et 104 filles dans le premier centre, et 87 garçons et 32 filles dans le second centre, calculer la moyenne générale des garçons puis celle des filles. Le président du jury en déduit que les garçons ont eu de meilleurs résultats que les filles. Est-ce vrai ?

Exercice 3

On a relevé la température moyenne chaque mois dans une ville :

Mois	Janv.	Fév.	Mars	Avr.	Mai	Juin	Jui.	Sept.	Oct.	Nov.	Déc.
Température	5	8	10	17	22	26	31	33	28	20	10

Déterminer la médiane, les quartiles, la moyenne et la variance de cette série statistique.

Exercice 4

On considère la série statistique suivante :

Classes	4	5	3.8	4.1	5.2	2	4.5	4.1

1) Déterminer la médiane de cette série.

2) Calculer la moyenne et la médiane de cette série statistique.

Exercice 5

1) Compléter le tableau ci-dessous qui donne la distribution des salaires mensuels bruts des 100 salariés d'une entreprise.

Salaires en euros	1500	1600	1900	2400	2700	3200	5000
Effectifs	30	25	15	12	8	6	4
Fréquences							

2) a) Donner le montant du salaire mensuel brut médian.

b) Calculer le pourcentage de la masse salariale totale perçue par les 10% des salariés les mieux rémunérés.

3) a) Calculer le montant du salaire mensuel brut moyen.

b) Calculer le pourcentage des salariés dont le salaire mensuel brut est compris dans l'intervalle $[\overline{x}-2\sigma;\overline{x}+2\sigma]$.

Exercice 6

Le tableau suivant donne le montant des salaires annuels exprimés en milliers d'euros d'une petite entreprise.

Salaires	16	18	20	25	30	40
Nombre de salariés	6	9	10	8	5	2

1) Déterminer la médiane, le premier et le troisième quartiles. Interpréter ces résultats et les traduire à l'aide d'un diagramme en boîte.

2) Calculer le montant en euros du salaire moyen annuel de cette entreprise.

3) A l'aide de la calculatrice, donner la valeur arrondie à l'euro près de l'écart-type s.

Exercice 7

Une entreprise de produits alimentaires fabrique et distribue une marque de café dans des sachets de 250 grammes. On suppose que le poids du sachet vide est négligeable.

La machine utilisée pour remplir les sachets est contrôlée selon la procédure suivante.

À chaque heure, un échantillon aléatoire de 40 sachets est prélevé dans la production ; on mesure la masse de chaque sachet et on calcule la masse moyenne $\overline{x}$ de l'échantillon.

Un réglage de la machine est nécessaire si l'un des critères suivants n'est pas vérifié :

- les 40 sachets ont une masse supérieure ou égale à 245 grammes ;
- 50% au moins des sachets ont une masse en grammes comprise dans l'intervalle $[248;252]$;
- 95% au moins des sachets ont une masse en grammes comprise dans l'intervalle $[\overline{x}-2\sigma;\overline{x}+2\sigma]$.

Au cours de la production, l'échantillon suivant a été prélevé.

248	256	253	246	252	250	248	253	248	251
255	252	252	254	250	250	250	251	250	252
256	252	251	247	251	249	253	250	251	245
250	252	247	249	250	249	249	249	254	249

1) Représenter la dispersion de cette série à l'aide d'un diagramme en boîte.

2) Faut-il effectuer un réglage de la machine ?