Chapitre 7

Statistiques

L'objectif de ce chapitre est d'étudier la moyenne, la médiane et l'étendue de séries statistiques et d'introduire la notion de moyenne pondérée et les indicateurs de dispersion que sont l'écart interquartile et l'écart type.

Le saviez-vous ?

Pour alerter les pouvoirs publics sur une épidémie, un mal du siècle, un comportement à risque, ..., les médecins s'appuient sur des statistiques. Mais ils s'en servent également pour juger de l'efficacité ou non d'un médicament.
Au XIX^{e} les statistiques et les représentations graphiques, dites à crête de coq, produites par l'infirmière F. Nightingale ont permis de diminuer de façon significative la mortalité des soldats britanniques lors de la guerre de Crimée (1853-1856).

1. Vocabulaire

Définitions

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Exemple :

- *On peut s'intéresser à une classe (population), comportant des élèves (individus) et observer leur nombre de frères et sœurs (caractère) qui peuvent être 0, 1, 2, ... (modalités), ces données formant alors une série statistique quantitative discrète.*
- *On peut s'intéresser à une chaîne d'usine produisant des bras de suspension pour voiture (population), et observer sur chaque pièce (individu) ses dimensions exactes (caractère) qui peuvent varier entre 500 et 750 mm (modalités), ces données formant alors une série statistique quantitative continue.*

— *On peut s'intéresser à la population française (population dont on prendra un échantillon) comportant des individus (individus) et estimer leur intention de vote (caractère) pouvant être n'importe lequel des candidats se présentant (modalités), ces données formant alors une série statistique qualitative.*

Définitions

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Souvent, il sera nécessaire de résumer les séries de valeurs : on produit alors *des* statistiques. Tout résumé met en évidence certaines caractéristiques de la série mais engendre une *perte d'information*, toutes les données n'étant plus accessibles.
Le résumé peut être un graphique : *diagramme en bâtons*, l'*histogramme* (pour des séries rangées en classes), une courbe de fréquences cumulées décroissantes, ...
Mais ce résumé peut aussi être numérique dans le cas d'une série statistique quantitative. Ces résumés numériques sont de deux types : les mesures centrales et les mesures de dispersion.

2. Indicateurs de tendance centrale

Ils visent à résumer la série par une seule valeur qu'on espère représentative de toutes les valeurs de la série.

a) Moyennes

Moyenne arithmétique

. . . . . .

. . . . . .

. . . . . .

Remarques :

— La somme de toutes les valeurs de la série est inchangée si on remplace chaque valeur par $\overline{x}$.

— On note parfois : $x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i$.

Moyenne pondérée

. . . . . .

. . . . . .

. . . . . .

. . . . . .

L'effectif total noté N est la somme de tous les effectifs : $N = n_1 + n_2 + \cdots + n_n$.

La moyenne a des avantages calculatoires : si l'on connaît les moyennes et les effectifs de deux séries, on peut obtenir la moyenne de la série constituée de l'agrégation de ces deux séries. Elle a le défaut d'être très sensible aux valeurs extrêmes.

Linéarité de la moyenne

Soient a et b deux réels. Si la série de valeurs (x_i) a pour moyenne m, alors la série de valeurs $(a \times x_i + b)$ aura pour moyenne $M = a \times m + b$.

Exemple : soient la série de notes suivantes : 12 ; 10,5 ; 13 ; 8,5. Sa moyenne est 11. Si à toutes les notes on applique la transformation suivante $1,25 \times note + 0,75$.
La nouvelle série de notes est : 15,75 ; 13,875 ; 17 ; 11,375 et l'on obtient comme moyenne 14,5 : valeur qui vérifie $1,25 \times 11 + 0,75$.

b) Médiane

Médiane

.....

.....

.....

Remarques :

- Rappel : mathématiquement « inférieur » et « supérieur » signifient, en français, « inférieur ou égal » et « supérieur ou égal ».
- On admettra qu'un tel nombre existe toujours.
- Plusieurs valeurs peuvent parfois convenir pour la médiane.
- La médiane partage la série en deux sous-séries ayant *quasiment* le même effectif ; *quasiment* car si plusieurs valeurs de la série sont égales à la médiane, les données inférieures à la médiane et les données supérieures à la médiane ne seront pas forcément en nombre égal.
- Il faut comprendre la médiane comme « la valeur du milieu ».

Propriétés

Soit une série statistique quantitative comportant n données : $S = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ telles que $x_1 \leq x_2 \leq \ldots \leq x_n$.

- ▷ Si n est impair, le $\frac{n+1}{2}^{\text{ième}}$ élément de la série est la médiane : $Me = x_{\frac{n+1}{2}}$
- ▷ Si n est pair, tout nombre compris entre le $\frac{n}{2}^{\text{ième}}$ élément de la série et le suivant est **une** médiane ; dans la pratique on prend la moyenne des deux données centrales de la série :

$$Me = \frac{\left(\frac{n}{2}\right)^{\text{ième}} + \left(\frac{n}{2}+1\right)^{\text{ième}}}{2}$$

La médiane a l'avantage de ne pas être influencée par les valeurs extrêmes. Elle n'a aucun avantage pratique dans les calculs, puisque pour connaître la médiane d'une série constituée de l'agrégation de deux séries, il faut nécessairement ré-ordonner la nouvelle série pour trouver sa médiane, qui n'aura pas de lien avec les deux médianes des deux séries initiales.

3. Indicateurs de tendance non centrale

Ils visent à indiquer comment les données de la série statistique sont dispersées par rapport aux mesures centrales.

a) Quartiles

Définition

.....

.....

.....

.....

.....

.....

.....

Comme pour la médiane, selon le nombre n de données dans la série, il y a parfois plusieurs possibilités pour Q_1 et Q_3 et parfois une seule, selon que n est ou n'est pas multiple de 4, ce qui peut compliquer leur recherche.
On convient de prendre systématiquement comme premier et troisième quartiles les nombres suivants :

Propriétés

Soit une série statistique quantitative comportant n données : $S = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ telles que $x_1 \leq x_2 \leq \ldots \leq x_n$. Alors :
- ▷ La donnée de rang $\frac{1}{4}n$, arrondi éventuellement au supérieur, convient toujours comme premier quartile.
- ▷ La donnée de rang $\frac{3}{4}n$, arrondi éventuellement au supérieur, convient toujours comme troisième quartile.

On l'admettra.

Exemple :
s'il y a $n = 29$ données dans la série, rangées dans l'ordre croissant :
- ∘ $\frac{1}{4} \times 29 = 7,25 \approx 8$ *donc la huitième donnée de la série convient comme premier quartile, soit 9 ;*
- ∘ $\frac{3}{4} \times 29 = 21,75 \approx 22$ *donc la vingt-deuxième donnée de la série convient comme troisième quartile, soit 11.*

x_i	8	9	10	11	12	13
n_i	3	5	7	7	4	3
e.c.c.	3	8	15	22	26	29

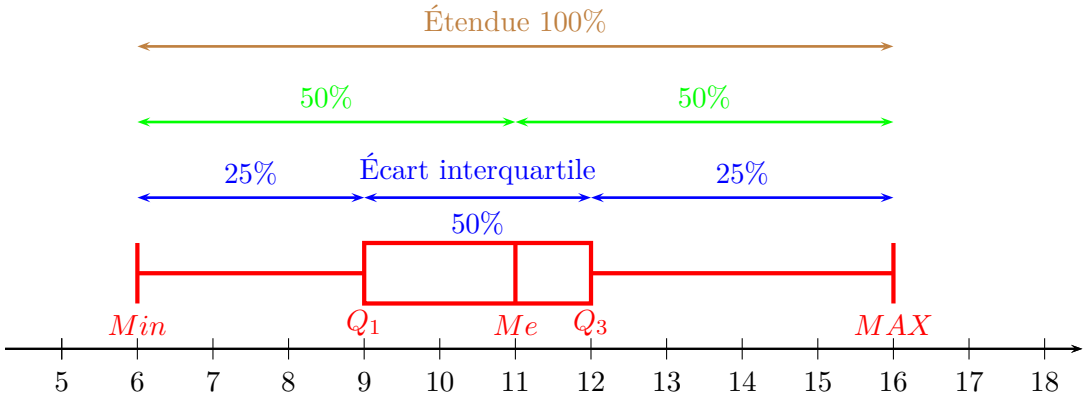
S'il y a $n = 64$ données dans la série, rangées dans l'ordre croissant :
- ∘ $\frac{1}{4} \times 64 = 16$ *donc la seizième donnée de la série convient comme premier quartile, soit 9 ;*
- ∘ $\frac{3}{4} \times 64 = 48$ *donc la quarante huitième donnée de la série convient comme troisième quartile, soit 12.*

x_i	6	7	8	9	10	11	12	13	14	15	16
n_i	2	5	5	7	12	14	9	4	3	2	1
e.c.c.	2	7	12	19	31	45	54	58	61	63	64

Propriété

Soit une série statistique quantitative comportant n données : $S = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ avec $n \geq 5$.
On ne change pas les quartiles et la médiane si on remplace x_1 par n'importe quel nombre de l'intervalle $]-\infty\,;\,x_1]$ et x_n par n'importe quel nombre de l'intervalle $[x_n\,;\,+\infty[$.

Comme on ne change pas le nombre de valeurs de la série, il y en aura toujours autant inférieures et supérieures à Q_1, Me et Q_3.



4. Indicateurs de dispersion

Valeurs extrêmes

. . . . . .

. . . . . .

. . . . . .

Écart interquartile

. . . . . .

. . . . . .

. . . . . .

Écart type

. . . . . .

. . . . . .

. . . . . .

. . . . . .

5. Comparaison de deux séries statistiques

Souvent, une série statistique est résumée par un couple associant un indicateur de tendance centrale à un indicateur de dispersion ; les couples utilisés sont moyenne et écart-type (m ; s) ou médiane et écart interquartile (Me ; $Q_3 - Q_1$). En comparant les indicateurs de tendance centrale des séries on peut déterminer la série qui semble être la plus forte (ou la plus faible) puis en comparant les indicateurs de dispersion celle qui est la plus régulière (ou la plus irrégulière).

6. Caisse à outils

Déterminer la moyenne et l'écart-type d'une série statistique $\Longrightarrow$ suivant la présentation des données, on doit tenir compte des valeurs et des effectifs ; la moyenne m (ou $\overline{x}$) est la division de la somme de toutes les valeurs par l'effectif total et l'écart-type s (ou σ) est la moyenne des écarts de toutes les valeurs par rapport à la moyenne. L'utilisation de l'application Statistiques de la calculatrice évitera des calculs fastidieux. Pour rappel les formules sont :

$$m = \frac{n_1.x_1 + n_2.x_2 + \ldots + n_n.x_n}{N} \qquad s = \sqrt{V} \text{ avec } V = \frac{n_1(x_1 - m)^2 + n_2(x_2 - m)^2 + \cdots + n_n(x_n - m)^2}{N}$$

Application :
Série 1 : une étude sur le nombre d'employés dans les commerces du centre d'une petite ville a donné les résultats suivants :

Nombre d'employés	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Effectif	*11*	*18*	*20*	*24*	*16*	*14*	*11*	*6*

Série 2 : une étude sur la durée de vie en heures de 200 ampoules électriques a donné les résultats suivants :

Durée de vie en centaine d'heures	*[10 ; 12[*	*[12 ; 14[*	*[14 ; 15[*	*[15 ; 16[*	*[16 ; 20[*
Effectif	*28*	*46*	*65*	*32*	*29*

pour les séries 1 et 2, déterminer :

1. *la moyenne ;*
2. *l'écart type.*

Déterminer la médiane et les quartiles d'une série statistique $\Longrightarrow$ les valeurs de la série doivent être triées dans l'ordre croissant, ensuite à partir de leur définition on recherche la position dans la série de valeurs de chacun des paramètres.
Médiane, si N est **impair**, la médiane occupe la position $\frac{N+1}{2}$ dans la série ; si N est **pair**, la médiane est la demi-somme des valeurs centrées $\frac{N}{2}$ième et $\frac{N+1}{2}$ième valeurs.
Quartiles, Q_1 premier quartile, c'est la plus petite donnée de la série telle qu'au moins un quart des données de la liste (25 %) sont **inférieures** ou **égales** à Q_1. Q_3 troisième quartile, c'est la plus petite donnée de la série telle qu'au moins trois quarts des données de la liste (75 %) sont **inférieures** ou **égales** à Q_3.
Min et *Max* valeurs extrêmes de la série respectivement la plus petite et la plus grande.

Effectif total pair : $N = 10$

n	1	2	3	4	5	6	7	8	9	10
x	12	12	13	14	14	15	15	15	16	16
			Q_1		Me			Q_3		

Moyenne : $\frac{12+12+\ldots+16}{10} = 14,2$
Médiane : $\frac{14+15}{2} = 14,5$
Position Q_1 : $\frac{10}{4} = 2,5 \quad \rightarrow 3^{\text{ième}}$ valeur donc $Q_1 = 13$
Position Q_3 : $\frac{10\times3}{4} = 7,5 \rightarrow 8^{\text{ième}}$ valeur donc $Q_3 = 15$

x	12	13	14	15	16
Eff.	2	1	2	3	2
Eff. c.	2	3	5	8	10

Effectif total impair : $N = 11$

n	1	2	3	4	5	6	7	8	9	10	11
x	12	12	13	14	14	15	15	15	16	16	17
			Q_1			Me			Q_3		

Moyenne : $\frac{12+12+\ldots+17}{11} \approx 14,45$
Médiane : 15
Position Q_1 : $\frac{11}{4} = 2,75 \rightarrow 3^{\text{ième}}$ valeur donc $Q_1 = 13$
Position Q_3 : $\frac{11\times3}{4} = 8,25 \rightarrow 9^{\text{ième}}$ valeur donc $Q_3 = 16$

x	12	13	14	15	16	17
Eff.	2	1	2	3	2	1
Eff. c.	2	3	5	8	10	11

Application : représenter les diagrammes en boîte correspondant aux deux séries suivantes :

Valeurs	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Effectifs série 1	*0*	*9*	*12*	*11*	*14*	*19*	*23*	*24*	*21*	*16*	*12*	*8*
Effectifs série 2	*11*	*10*	*11*	*13*	*17*	*20*	*22*	*23*	*20*	*11*	*6*	*0*

Pour choisir entre les couples $(Moy\ ;\ \sigma)$ *et* $(Me\ ;\ \Delta_Q) \Longrightarrow$ le couple moyenne / écart-type prend en compte toutes les valeurs de la série donc il est influencé par les valeurs extrêmes. Le couple médiane / écart interquartile est déterminé par le nombre et non la taille des valeurs donc il ne sera pas influencé par les valeurs extrêmes parfois trompeuses. Si l'on doit regrouper les résultats de plusieurs sous-groupes pour mener à bien une étude, on retiendra le couple $(Moy\ ;\ \sigma)$ qui se prête mieux aux calculs.

Application : choisir le couple le mieux adapté pour résumer les séries proposées ci-après :

Série 1	*2*	*2*	*3*	*5*	*5*	*18*	*20*
Série 2	*8*	*10*	*12*	*12*	*12*	*14*	*16*
Série 3	*4*	*10*	*10*	*12*	*14*	*14*	*20*

Pour comparer ou étudier deux séries de valeurs $\Longrightarrow$ on peut tracer leurs diagrammes en boîte sur un même graphique puis on qualifie la position des valeurs par leur importance (position centrale : Me), leur amplitude (dispersion : Δ_Q). Si on utilise l'autre couple, pour des séries de même nature plus la valeur de l'écart-type est importante plus la série est dispersée.

Application : comparer les deux séries proposées ci-après :

	Min	Q_1	Me	Q_3	Max
Série 1	*14.5*	*16*	*16.5*	*18*	*19.5*
Série 2	*12*	*15*	*23*	*28*	*32*

7. Algorithmes

Les valeurs de la série sont stockées dans la liste *serie*.
La première valeur dans la liste occupe le rang 0 donc il faudra y être vigilant notamment dans la recherche de la médiane ou des quartiles. (décalage de 1 unité de la position de la valeur)

a) Indicateurs de position centrale

Calcul de la moyenne d'une série de valeurs

Pour déterminer la moyenne d'une série de valeurs, on doit ajouter toutes les valeurs de la série puis diviser cette somme par l'effectif total. Les commandes ci-contre peuvent être exploitées au sein d'une fonction. La commande **len** donne le nombre de valeurs dans la liste donc l'effectif total.

$N \leftarrow$ effectif total
$T \leftarrow 0$
pour *i variant de 0 à N* **faire**
  $T \leftarrow T + valeur[i]$
fin
$m \leftarrow T/N$
Afficher m

```
N = len(serie)
T = 0
for i in range(N) :
        T = T + serie[i]
m = T / N
print(m)
```

Pour être plus efficace encore, on utilise la commande **sum** qui effectue la somme des valeurs d'une liste.

```
m = sum(serie) / len(serie)
print(m)
```

Calcul de la médiane d'une série de valeurs

Pour déterminer la médiane d'une série de valeurs, on doit trier les valeurs dans l'ordre croissant, chose réalisée par la commande **sorted** sous Python. Ensuite suivant la parité de l'effectif total, on recherche la valeur partageant la série en deux parties de même effectif ou l'on calcule la moyenne des valeurs centrales.

Trier dans l'ordre croissant les valeurs
$N \leftarrow$ effectif total
si *N est impair* **alors**
  $Me \leftarrow valeurtriee[(N+1)/2]$
sinon
  $Me \leftarrow \frac{valeurtriee[N/2]+valeurtriee[(N+1)/2]}{2}$
fin
Afficher Me

```
trie = sorted(serie)
N = len(trie)
if N % 2 == 1 :
        Me = trie[N // 2]
else :
        Me = (trie[N // 2 - 1] +
          trie[N // 2]) / 2
print(Me)
```

b) Indicateurs de dispersion

Calcul de l'étendue d'une série de valeurs

Pour déterminer l'étendue d'une série, on effectue la différence entre les valeurs extrêmes, la plus grande - la plus petite. En utilisant les commandes **max** et **min**, l'écriture du programme peut se limiter à une unique ligne.

$E \leftarrow MAX(valeurs) - min(valeurs)$
Afficher E

```
print("Etendue = ",max(serie)-min(serie))
```

Calcul de l'écart type d'une série de valeurs

Pour déterminer l'écart type d'une série, on effectue la moyenne des écarts positifs des valeurs par rapport à la moyenne. On doit penser à importer la collection math pour pouvoir effectuer le calcul de la racine carrée.

$N \leftarrow$ effectif total
$m \leftarrow$ moyenne(serie)
pour *i variant de 1 à N* **faire**
  $ecarts[i] \leftarrow (serie[i] - m)^2$
fin
$s \leftarrow \sqrt{\frac{somme(ecarts)}{N}}$
Afficher s

```
from math import*

def moyenne(maliste):
    return sum(maliste)/len(maliste)

N = len(serie)
m = moyenne(serie)
ecarts = [(serie[i]-m)**2 for i in range(N)]
s = sqrt(sum(ecarts)/N)

print("Ecart type, s = ",s)
```

Calcul de l'écart interquartile d'une série de valeurs

Pour déterminer l'écart interquartile, au préalable, on doit connaître le premier et troisième quartile. Ensuite il ne reste plus qu'à faire la différence entre le troisième et le premier quartile.

Trier dans l'ordre croissant les valeurs
$N \leftarrow$ effectif total
$Q1 \leftarrow valeurtriee[arrondi.sup(N/4)]$
$Q3 \leftarrow valeurtriee[arrondi.sup(3N/4)]$
$\Delta Q \leftarrow Q3 - Q1$ Afficher ΔQ

```
def q1(maliste):
    trie = sorted(maliste)
    N = len(trie)
    return trie[N//4]

def q3(maliste):
    trie = sorted(maliste)
    N = len(trie)
    return trie[N*3//4]

print("Q1 = ",q1(serie)," et Q3 = ",
q3(serie))
print("Ecart interquartile : ",q3(serie)-
q1(serie))
```

c) Proportion d'éléments dans un intervalle

Calcul de la proportion de valeurs appartenant à l'intervalle $[m-2s\ ;\ m+2s]$

Pour déterminer la proportion de valeurs d'une série appartenant à l'intervalle $[m-2s\ ;\ m+2s]$, on doit connaître la moyenne et l'écart type de la série. Ensuite, en utilisant un test et un compteur incrémenté en fonction du résultat du test, on compte le nombre de valeurs présentes dans l'intervalle. Pour en obtenir la proportion, ne reste plus qu'à diviser ce nombre par l'effectif total de la série.

$m \leftarrow$ moyenne(serie)
$s \leftarrow$ ecartype(serie)
$nbi \leftarrow 0$
pour *toutes les valeurs de (serie)* **faire**
  si $m-2*s<val$ *et* $val<m+2*s$ **alors**
    $nbi \leftarrow nbi+1$
  fin
fin
$frequence \leftarrow \frac{nbi}{N}$
Afficher $m, s, frequence$

```
from math import*

def moyenne(maliste):
  return sum(maliste)/len(maliste)

def ecartype(maliste):
  N = len(maliste)
  m = moyenne(maliste)
  ecarts = [(maliste[i]-m)**2 for i in range(N)]
  s = sqrt(sum(ecarts)/N)
  return s

def propor(maliste):
  m = moyenne(maliste)
  s = ecartype(maliste)
  nbi = 0
  for t in maliste :
    if m-2*s < t and t < m+2*s :
      nbi = nbi + 1
  frequence = nbi / len(maliste)
  return m, s, frequence

m,s,p = propor(serie)
print("Moyenne , m =",m)
print("Ecart type, s = ",s)
print("Intervalle : [",m-2*s," ; ",m+2*s,"]")
print("Proportion, p = ",p)
```

Calcul de la proportion de valeurs appartenant à l'intervalle $[m-2s\ ;\ m+2s]$ avec représentation graphique

En plus des instructions précédentes, on doit importer un module de traçage et les lignes suivantes :

```
import matplotlib.pyplot as plt

Liste_X = range(1, len(serie)+1)
Liste_Y = serie
plt.plot(Liste_X,Liste_Y,"b.")
plt.plot((0,len(serie)+1),(m,m),"g")
plt.plot((0,len(serie)+1),(m-2*s,m-2*s),"r")
plt.plot((0,len(serie)+1),(m+2*s,m+2*s),"r")
plt.show()
```

8. Évaluations

Devoir en temps libre n° 7 : Statistiques

Il est rappelé que la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'apprécitaion des copies. Le barème est donné à titre indicatif. Le sujet sera rendu avec la copie.

Exercice n°1 : ***Quelque soit la présentation des données ...***

Série 1

17.8	18.3	18.5	17.9	18.5	18.2	18.1	18.1	18.3	17.6
17.2	18	17.6	17.9	18.3	18.5	17.7	18.2	18	18.4

Série 2

x_i	5	7	8.5	9	10	11	12	13	15	18	20
n_i	2	4	5	7	11	12	9	4	3	2	1

1. Avec la série 1
   a) Tester les programmes proposés dans le cours pour tous les indicateurs.
   b) Vérifier les résultats avec la calculatrice.
2. Avec la série 2
   a) Déterminer avec la calculatrice tous les indicateurs.
   b) Écrire et tester des programmes sous Python permettant d'obtenir la moyenne, l'écart type, la médiane et les quartiles 1 et 3.
      A minima, il est attendu une feuille de calculs sous tableur permettant d'obtenir les résultats souhaités. La version experte serait l'écriture de fonctions donnant les indicateurs à partir de deux listes l'une contenant les valeurs, l'autre les effectifs.
   c) Vérifier vos résultats.

Exercice n°2 : ***Déterminer des effectifs***

Reproduire et compléter le tableau proposé ci-après afin que :
- la moyenne soit égale à 2 ;
- la médiane soit égale à 1 ;
- les quartiles soient $Q_1 = 0$ et $Q_3 = 3$.

Valeur	-1	0	1	3	5	Total
Effectif						25

Proposer deux solutions différentes, pour information il en existe 13 ...