Chapitre XII : Les statistiques descriptives

I - Série statistique

a) Effectifs et fréquences

Dans une **population** choisie, on étudie un **caractère** particulier. Celui-ci peut prendre différentes valeurs selon les **individus**.

Si le caractère étudié prend des valeurs numériques, on dit qu'il est **quantitatif**. Sinon, il est **qualitatif**.
On étudiera uniquement des situation avec un caractère **quantitatif discret**, ce qui signifie que les valeurs sont isolées et peuvent être énumérées (contrairement à un caractère **caractère quantitatif continu** où les valeurs seraient dans un intervalle, comme pour des distances, des données de temps, ...).

Exemple :
Au cours d'une enquête portant sur les bébés nés en 2019, on s'intéresse :

- à la couleur de leurs yeux : caractère qualitatif (marron, bleu, ...) ;
- au nombre de fois où ils pleurent par jour : caractère quantitatif discret $(1\,;\,2\,;\,3\,;\,\ldots)$;
- à leur taille en centimètres : caractère quantitatif continu $([40\,;\,45[\,;\,[45\,;\,50[\,;\,\ldots)$.

Définition :

- L'**effectif total** est le nombre d'individus de la population.
- L'**effectif d'une valeur** est le nombre d'individus de la population dont le caractère prend cette valeur. Par moment celui-ci n'apparaitra pas et on considèrera que chaque valeur a pour effectif 1 (y compris celles qui sont répétées, sauf qu'on les écrira plusieurs fois).
- L'**effectif cumulé croissant** (E.C.C.) d'une valeur est le nombre d'individus dont le caractère a une valeur inférieure ou égale à celle-ci.

Remarque : L'étude de données statistiques se fait très souvent sur un nombre important de valeurs. Les outils numériques comme le **tableur** et les **algorithmes** permettent d'effectuer les calculs sur des données très importantes et prennent donc une place importante dans ce chapitre.

Exemple : Des amateurs de tir à l'arc ont effectué un parcours comportant 24 cibles. Le tableau ci-dessous (feuille de calcul d'un tableur) associe au nombre de cibles touchées, appelé « score », l'effectif de tireurs correspondant.

	A	B	C	D	E	F	G	H	I	J	K	L
1	scores	7	8	10	11	12	13	14	15	16	18	Total
2	effectifs	2	1	3	3	5	4	7	3	2	1	31
3	E.C.C.	2	3	6	9	14	18	25	28	30	31	

Dans cet exemple, la population est l'ensemble des tirs effectués et le caractère choisi est l'ensemble des tirs réussis.
L'effectif total correspond au nombre total de tireurs. Il est égal à .....

Dans la feuille de calcul, la formule à saisir en L2 pour obtenir cette valeur est : .....

L'effectif cumulé croissant de la valeur 14 vaut 25. Cela signifie que 25 tireurs ont touché 14 cibles ou moins (on peut aussi dire « au plus 14 cibles »).
Dans la feuille de calcul, en B3 on écrit ..... puis en C3 on écrit .....
En recopiant (ou tirant) cette formule vers la droite jusqu'en K3, on complète alors entièrement la ligne 3.

En D3, cela nous donne la formule ....., ainsi de suite...

Définition : Dans une série statistique, la **fréquence** f d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total : $f = \dfrac{\text{effectif de la valeur}}{\text{effectif total}}$.
À l'aide de la fréquence f d'une valeur et de l'effectif total, on retrouve l'effectif par le calcul :
effectif de la valeur $= f \times$ effectif total.
La **fréquence cumulée croissante** (F.C.C.) d'une valeur est la fréquence d'individus dont le caractère a une valeur inférieure ou égale à celle-ci.

Ajoutons deux lignes dans la feuille de calcul de l'exemple 1 :

	A	B	C	D	E	F	G	H	I	J	K	L
1	scores	7	8	10	11	12	13	14	15	16	18	Total
2	effectifs	2	1	3	3	5	4	7	3	2	1	31
3	E.C.C.	2	3	6	9	14	18	25	28	30	31	
4	Fréquence	0,065	0,032	0,097	0,097	0,161	0,129	0,226	0,097	0,065	0,032	
5	F.C.C.	0,065	0,097	0,194	0,290	0,452	0,581	0,806	0,903	0,968	1	

En B4, on souhaite effectuer le calcul $\frac{2}{31}$ qui vaut environ 0,065.
Dans la cellule B4, on peut alors écrire la formule = B2/L2.
En recopiant cette formule vers la droite, on n'obtient pas la formule attendue.

En C4, on a la formule ..... alors qu'on aurait voulu avoir .....

Il convient alors de fixer la lettre L de L2, ce qui correspond à la colonne L. Cela se fait à l'aide du symbole \$.

Donc en B4 on doit écrire ..... et on peut recopier cette formule vers la droite.

b) Traitement sous la forme d'un algorithme

Pour effectuer un traitement algorithmique de ces données, il convient d'utiliser un type de données particulier : les listes. C'est d'ailleurs ce que nous ferons avec la calculatrice.
En python, on définit une liste en utilisant des **crochets** et en séparant les valeurs par des **virgules** (je rappelle que le séparateur des décimaux est le point). Par exemple :

```
scores=[7, 8, 10, 11, 12, 13, 14, 15, 16, 18]
effectifs=[2, 1, 3, 3, 5, 4, 7, 3, 2, 1]
```

On obtient très facilement le nombre de valeurs de la liste : `len(scores)` (qui vaut 10 ici), `len` étant une contraction de « length » signifiant « longueur ».

Les valeurs présentes dans la liste `scores` peuvent être obtenues par leur position dans la liste **sachant que la première position est toujours 0**. Pour cela on utilise de nouveau des crochets : `scores[0]` vaut 7, `scores[5]` vaut 13 (c'est la 6ème valeur de la liste).

Le parcours d'une liste se fait très facilement à l'aide d'une boucle `for` même s'il existe deux méthodes :

- `for elt in scores :`

  où `elt` prendra successivement les valeurs de la liste (7 au premier passage, 8 au deuxième, ...).

- `for i in range(len(scores)):`

  La liste `scores` ayant pour longueur 10, `i` prendra successivement les valeurs allant de 0 à 9 (rappelez-vous le travail effectué sur les boucles).

Avec toutes ces informations, on peut, à partir des listes `scores` et `effectifs`, calculer l'effectif total. Avant la boucle, cet effectif total vaut 0 et à chaque passage, on doit ajouter l'effectif à l'effectif total intermédiaire.

```
scores=[7, 8, 10, 11, 12, 13, 14, 15, 16, 18]
effectifs=[2, 1, 3, 3, 5, 4, 7, 3, 2, 1]
effTotal=0
for valeur in effectifs:
   effTotal=effTotal+valeur
```

On peut également construire une nouvelle liste nommée `ECC`. Avant la boucle, cette liste contiendra le premier effectif (comme pour le tableur) et à chaque passage dans la boucle, on lui ajoutera une valeur (avec la méthode `append`) égale à son dernier élément ajouté au nouvel effectif :

```
ECC=[effectifs[0]]
for i in range(1, len(effectifs)):
   ECC.append(ECC[i-1]+effectifs[i])
```

On peut également créer la liste des fréquences en python (la liste des fréquences cumulées croissantes s'effectuant comme pour les effectifs cumulés croissants), toujours à l'aide d'une boucle `for` (bien sûr, on utilise la variable `effTotal` déterminée précédemment) :

```
frequences=[]
for valeur in effectifs:
   frequences.append(valeur/effTotal)
```

II - Médiane et écart interquartile

a) Des paramètres de position : la médiane et les quartiles

Définition : La médiane d'une série statistique **ordonnée** partage cette série en deux parties de telle sorte que :

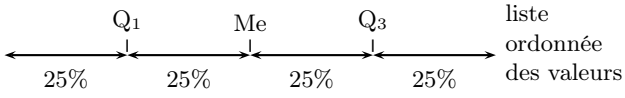
- au moins la moitié des valeurs sont inférieures ou égales à la médiane ;
- au moins la moitié des valeurs sont supérieures ou égales à la médiane.

Méthode pour une série au caractère quantitatif discret :
Si la série contient n valeurs **rangées dans l'ordre croissant** :

- si n est impair, on prend la $\left(\frac{n}{2}+\frac{1}{2}\right)$ ème valeur pour médiane.
- si n est pair, on prend pour médiane la moyenne entre la $\frac{n}{2}$ ème et la $\frac{n}{2}+1$ ème valeur.

Exemple : Dans l'exemple de la partie I, l'effectif total vaut 31 (nombre impair).
$\frac{31}{2} = 15{,}5$. Donc on prend pour médiane la $16^{\text{ème}}$ valeur (qu'on peut déterminer à l'aide des effectifs cumulés croissants).
Donc la médiane vaut 13.

Définition des quartiles :
Le **premier quartile** Q_1 de la série est la plus petite donnée telle qu'au moins 25% des valeurs de la série lui soient inférieures ou égales.
Le **troisième quartile** Q_3 de la série est la plus petite donnée telle qu'au moins 75% des valeurs de la série lui soient inférieures ou égales.



Exemple :

Dans l'exemple de la partie I, $\frac{31}{4} = 7{,}75$. Ainsi le premier quartile est la $8^{\text{ème}}$ valeur. Donc $Q_1 = 11$.
$3 \times \frac{31}{4} = 23{,}25$. Donc le troisième quartile est la $24^{\text{ème}}$ valeur. Donc $Q_3 = 14$.

b) Un paramètre de dispersion : l'écart interquartile

Définition : En troisième, vous avez découvert l'**étendue** d'une série statistique qui est la différence entre la plus grande et la plus petite valeur de cette série.

Exemple : Dans l'exemple de la partie I, l'étendue vaut $18 - 7 = 11$.

Définition : L'**écart interquartile** est la différence $\boldsymbol{Q_3 - Q_1}$ entre les troisième et premier quartiles.

Exemple : Dans l'exemple de la partie I, l'écart interquartile vaut $Q_3 - Q_1 = 14 - 11 = 3$.

Remarque : Ces deux paramètres peuvent servir à comparer des séries statistiques traitant d'un même thème sur deux populations différentes.
Plus l'écart interquartile est élevé, plus les données sont considérées comme étant **hétérogènes** ou **dispersées**.
Plus l'écart interquartile est faible, plus les données sont considérées comme **homogènes** (ou **non dispersées**).

L'étendue a un rôle similaire même si l'interprétation peut vite être faussée si la série de données contient une valeur extrême (très supérieure ou très inférieure aux autres).

Dans le manuel, vous avez un exemple page 272 d'une série contenant un nombre pair de valeur (valeurs données sous la forme d'une liste, sans effectif). Cela vous montre la méthode et **la rédaction** permettant de trouver la médiane et les quartiles.
Vous avez un autre exemple en haut de la page 273 (cas d'un effectif total impair et d'un tableau avec effectifs - c'est à vous de calculer les effectifs cumulés croissants pour pouvoir déterminer la médiane et les quartiles).