26 Statistiques univariées

ECG Maths Appliquées
Semestre 1

Un grand nombre de situations dans la vie courant amène à effectuer des enquêtes statistiques. Une enquête statistique est l'**observation d'une population** (population d'un pays, élèves d'une classe, animaux présents dans une réserve, ...) et à étudier une ou plusieurs grandeur(s) ou caractéristique(s) appelée(s) **caractère(s) observé(s)** (âge, langues parlées, note en maths à HEC, couleur des yeux, ...) dans cette population. Si les valeurs prises par le caractère sont des nombre réels, on parle de statistique quantitative ; dans le cas contraire, on parle de statistique qualitative. Dans ce chapitre, on ne s'intéresse qu'à l'étude des séries statistiques quantitatives.

A. Population, caractère et statistique

Définition 26.1

i. Étant donné un ensemble fini non vide Ω appelé **population** et dont les éléments sont appelés **individus**, on appelle **statistique quantitative** sur Ω toute application X de Ω dans $\mathbb{R}$. Le cardinal de Ω est appelé **effectif total de** la population, en général noté n.

ii. Si le cardinal de effectif total de est n et si les individus sont numérotés $\omega_1, \dots, \omega_n$, la série statistique associée à x est la suite $(x(\omega_2), \dots, x(\omega_n))$ des valeurs du caractère relevées dans la population.

Remarques

a. Même si la population est finie, il arrive fréquemment que l'on ne puisse pas étudier tous les individus de la population, soit parce que la population est trop importante et qu'une étude complète serait trop coûteuse ou trop longue, soit parce qu'une étude complète n'est pas réalisable (pour étudier la durée de vie des ampoules produites par une usine, est-il réaliste de les utiliser toutes jusqu'à extinction des feux ?). Dans ce cas, on est conduit à limiter l'étude à une partie de la population, appelée échantillon : on parle de **statistique inférentielle**. L'étude statistique ainsi réalisée ne donne alors pas une image réelle de la population, mais seulement une estimation. La théorie de l'estimation faisant l'objet d'un autre chapitre, nous n'allons pas insister dessus ici et considèrerons donc que l'étude statistique est menée sur l'ensemble de la population.

b. Une statistique quantitative permet donc d'étudier un caractère en relevant, pour chaque individu la valeur de son caractère (sa taille, son poids, le nombre de frères ou de sœurs,...).

Définition 26.2

i. On dit que la série statistique x est une **série discrète** si l'on peut considérer que x prend ses valeurs dans un ensemble fini ; dans le cas d'une série statistique discrète, si les valeurs $x_1, \dots, x_p$ que peut effectivement prendre x sont appelées **modalités** et si, pour tout $i \in [\![1, p]\!]$, on note n_i le nombre d'individus pour lesquels la valeur du caractère est x_i (appelé **effectif** de x_i), la série peut être notée $((x_i, n_i))_{1 \leqslant i \leqslant p}$.

ii. On dit que la série statistique x est une **série continue** si l'on considère que x prend ses valeurs dans un intervalle $]a, b[$ de $\mathbb{R}$; dans ce cas, si $(a_i)_{1 \leqslant i \leqslant p+1}$ est une suite strictement croissante tel que x prend ses valeurs dans $]a_1, a_{p+1}[$, les données peuvent être regroupées dans les intervalles $[x_1, x_2[$, $[x_2, x_3[$, ..., $[x_p, x_{p+1}[$, appelés **classes** de la série et si, pour tout $i \in [\![1, p]\!]$, on note n_i le nombre d'individus pour lesquels le valeur du caractère appartient à $[a_i, a_{i+1}[$ (appelé **effectif de la classe**), la série peut être notée $([a_i, a_{i+1}[, n_i)_{1 \leqslant i \leqslant p+1}$.

Remarques

a. Il arrive souvent que les valeurs possibles du caractère étudié soient en nombre fini (donc que la série soit discrète) mais que l'effectif total de la population ou le nombre total de

valeurs possibles soit trop important pour que les différentes valeurs relevées fournissent une information intéressante. C'est dans ce cas que l'on utilise en général des séries statistiques continues, les données étant groupées par intervalles, aussi appelés classes ; on considère alors que, dans chaque classe les valeurs sont réparties de manière uniforme. C'est le cas par exemple si l'on étudie la taille d'un individu, que l'on regroupe implicitement par classe en omettant en général les millimètres !

b. Une série statistique continue peut aussi être de la forme du moment que les intervalles sont deux à deux disjoints et que leur réunion contient l'ensemble des valeurs prises par le caractère dans la population étudiée. Cependant, par souci de simplicité, on n'envisagera dans la suite que des séries statistiques discrètes ou des séries statistiques continues de la forme

c. Dans la suite, quand on envisagera la série statistique discrète $((x_i, n_i))_{1\leqslant i\leqslant p}$, on considèrera que la suite est strictement croissante.

Définition 26.3

Soit $((x_i, n_i))_{1\leqslant i\leqslant p}$ une série statistique discrète.

i. Pour tout $i \in [\![1, p]\!]$, n_i est appelé effectif de x_i.

ii. L'entier $n = \sum_{i=1}^{p} n_i$ est appelé effectif total de la population.

iii. Pour tout $i \in [\![1, p]\!]$, le nombre $n_1 + n_2 + \cdots + n_i$ est appelé effectif cumulé en x_i ; c'est le nombre d'individus de la population pour lesquels la valeur du caractère étudié est inférieur ou égal à x_i.

iv. Pour tout $i \in [\![1, p]\!]$, le nombre $f_i = \frac{n_i}{n}$ est appelé fréquence de x_i ; c'est la proportion d'individus de la population pour lesquels la valeur du caractère étudié est égale à x_i.

v. Pour tout $i \in [\![1, p]\!]$, le nombre $f_1+f_2+\cdots+f_i$ est appelé fréquence cumulée en x_i ; c'est la proportion d'individus de la population pour lesquels la valeur du caractère étudié est inférieure ou égale à x_i.

On adopte des définitions analogues dans le cas d'une série statistique continue $([a_i, a_{i+1}[, n_i)_{1\leqslant i\leqslant p+1}$, en remplaçant le réel x_i par la classe $[_i, a_{i+1}[$.

Exemples 26.1 **a.** On a effectué une suite de 60 lancers successifs d'un dé à six faces et on a obtenu les résultats suivants :

Résultat	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
1	11	11	0,183	0,183
2	6	17	0,1	0,283
3	12	29	0,2	0,483
4	15	44	0,25	0,733
5	9	53	0,15	0,883
6	7	60	0,117	1

b. Un magazine a fait une étude sur la durée de vie des écrans de téléphone portable. Sur un échantillon de 340 ordinateurs étudiés, elle a obtenu les résultats suivants :

Résultat	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
$[0, 12[$	91	91	0,268	0,268
$[12, 18[$	40	131	0,118	0,385
$[18, 24[$	47	178	0,138	0,524
$[24, 36[$	82	260	0,241	0,765
$[36, 48[$	62	322	0,182	0,947
$[48, 60[$	18	340	0,053	1

c. Dans le tableau suivant, on s'intéresse à la répartition des nombres premiers compris entre 0 et 100 :

Résultat	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
$[0,10[$	4	4	0,16	0,16
$[10,20[$	4	8	0,16	0,32
$[20,30[$	2	10	0,08	0,4
$[30,40[$	2	12	0,08	0,48
$[40,50[$	3	15	0,12	0,6
$[50,60[$	2	17	0,08	0,68
$[60,70[$	2	19	0,08	0,76
$[70,80[$	3	22	0,12	0,88
$[80,90[$	2	24	0,08	0,96
$[90,100[$	1	25	0,04	1

Proposition 26.4

Si $(f_i)_{1\leqslant i\leqslant p}$ est la suite des fréquences d'une série statistique à p modalités (ou p classes), alors :

$$\sum_{im=1}^{p} f_i = 1$$

Définition 26.5

Soit $x = (x_i, n_i)_{1\leqslant i\leqslant p}$ une série statistique et (a, b) un couple de réels tel que $a \neq 0$.
La série statistique $y = (y_i, n_i)_{1\leqslant i\leqslant p}$ telle que, pour tout $i \in [\![1, p]\!]$, $y_i = ax_i+b$ est appelée transformation affine de x. On la note aussi $y = ax + b$.
On adopte une définition analogue dans le cas d'une série statistique continue.

B. Caractéristiques de position

En statistiques, l'objectif est souvent de résumer la série par un petit nombre de valeurs. La question est donc de savoir quelle(s) valeur(s) est (sont) susceptible(s) de résumer le mieux l'information disponible.

B.1. Moyenne

Définition 26.6

i. On appelle **moyenne** de la série statistique $x = ((x_i, n_i))_{1\leqslant i\leqslant p}$ le nombre

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \sum_{i=1}^{n} x_i f_i$$

où $n = \sum_{i=1}^{n} n_i$ est l'effectif total de la série.

ii. On appelle **moyenne** de la série statistique $x = ([a_i, a_{i+1}[, n_i)_{1\leqslant i\leqslant p+1}$ le nombre

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \sum_{i=1}^{n} x_i f_i$$

où $n = \sum_{i=1}^{n} n_i$ est l'effectif total de la série et $x_i = \frac{a_i+a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

Exemples 26.2 **a.** Dans la première série statistique de l'exemple 23.1 (lancers de dés), la moyenne est $3,43$.

b. Dans la deuxième série statistique de l'exemple 23.1 (lancers de dés), la moyenne est $24,03$.

c. Dans la troisième série statistique de l'exemple 23.1 (lancers de dés), la moyenne est $42,6$ (cela dit, il est douteux que la moyenne ait une réelle utilité dans ce cas).

Proposition 26.7

Si x est une série statistique de moyenne $\overline{x}$ et si $y = ax + b$ est une transformation affine de x, alors la moyenne $\overline{y}$ de y vérifie :

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

B.2. Médiane

Définition 26.8

Soit x une série statistique discrète dont les relevés (non forcément distincts) sont $x_1, \dots, x_n$, rangés dans l'ordre croissant.
On appelle médiane de x toute valeur m qui sépare la série en deux séries de même effectif de telle sorte qu'au moins la moitié soient inférieures ou égales à m et au moins la moitié soient supérieures ou égales à m.

Proposition 26.9

Soit x une série statistique discrète dont les relevés (non forcément distincts) sont $x_1, \dots, x_n$, rangés dans l'ordre croissant.

i. Si n est impair, il n'y a qu'une seule médiane et, en notant $n = 2k + 1$, la médiane est x_k.

ii. Si n est pair, deux cas sont possibles en notant $n = 2k$

- si $x_k = x_{k+1}$, alors x_k est l'unique médiane,
- si $x_k < x_{k+1}$, il y a une infinité de médianes : tous les réels appartenant à $[x_k, x_{k+1}[$; dans ce cas, on choisit par convention de dire que **la** médiane est $\frac{x_k + x_{k+1}}{2}$.

Exemples 26.3

a. Pour la première série statistique de l'exemple 23.1 (lancers de dés), la médiane est 4, car l'effectif total est égal à 60 et les $30^{ème}$ et $30^{ème}$ valeurs sont égales à 4.

b. La série statistique $(1, 1, 2, 2, 3, 3, 4, 6, 9)$ contient 9 relevés (nombre impair) et la cinquième valeur dans l'ordre croissant est 3, donc la médiane est 3.

c. La série statistique $(3, 3, 3, 2, 3, 1, 1, 2, 1, 3)$ contient 10 valeurs (nombre pair) et la cinquième valeur dans l'ordre croissant est 2 et la sixième est 3, donc la médiane est $\frac{2+3}{2} = 2,5$.

Remarque

On parle parfois aussi de médianes (au pluriel) pour désigner tout réel m tel que 50% des individus ont une valeur du caractère inférieure ou égale à m et 50% ont une valeur supérieure ou égale à m, autrement dit tel que :

$$\sum_{\substack{1 \leqslant i \leqslant p \\ x_i < m}} f_i < \frac{1}{2} \leqslant \sum_{\substack{1 \leqslant i \leqslant p \\ x_i \leqslant m}} f_i$$

Définition 26.10

Soit $x = ([a_i, a_{i+1}[, n_i)_{1 \leqslant i \leqslant p+1}$ une série statistique continue. Pour tout $i \in [\![1, p]\!]$, on note f_i la fréquence de $[a_i, a_{i+1}[$.
La ligne brisée joignant les points $M_0(a_0, 0), M_1(a_2, f_1), M_2(a_2, f_1 + f_2), \dots, M_i(a_i, f_1 + f_2 + \cdots f_i), \dots, M_p(a_{p+1}, 1)$ est appelée **polygone des fréquences cumulées** de la série x.

Définition 26.11

Soit $x = ([a_i, a_{i+1}[, n_i)_{1 \leqslant i \leqslant p+1}$ une série statistique continue.
On appelle **médiane** de x l'abscisse du point du polygone des fréquences cumulées dont l'ordonnée est égale à $\frac{1}{2}$.

Remarques

a. Dans le cas d'une série statistique continue (mais aussi dans le cas d'une série statistique discrète), on peut donc déterminer la médiane graphiquement, mais cela demande un représentation précise et on obtiendra le plus souvent une valeur approchée.

b. Pour déterminer la valeur précise de la médiane, on peut procéder ainsi :

- on commence par déterminer l'entier i tel que la médiane appartienne à la classe $[a_i, a_{i+1}[$, c'est-à-dire tel que :

$$f_1 + f_2 + \cdots + f_{i-1} < \frac{1}{2} \leqslant f_1 + f_2 + \cdots + f_{i-1} + f_i$$

- puis on détermine le point d'ordonnée $\frac{1}{2}$ sur la droite $(M_{i-1}M_i)$, dont l'équation est :

$$y = \frac{f_i}{a_{i+1} - a_i}(x - a_i) + \sum_{j=1}^{i-1} f_j$$

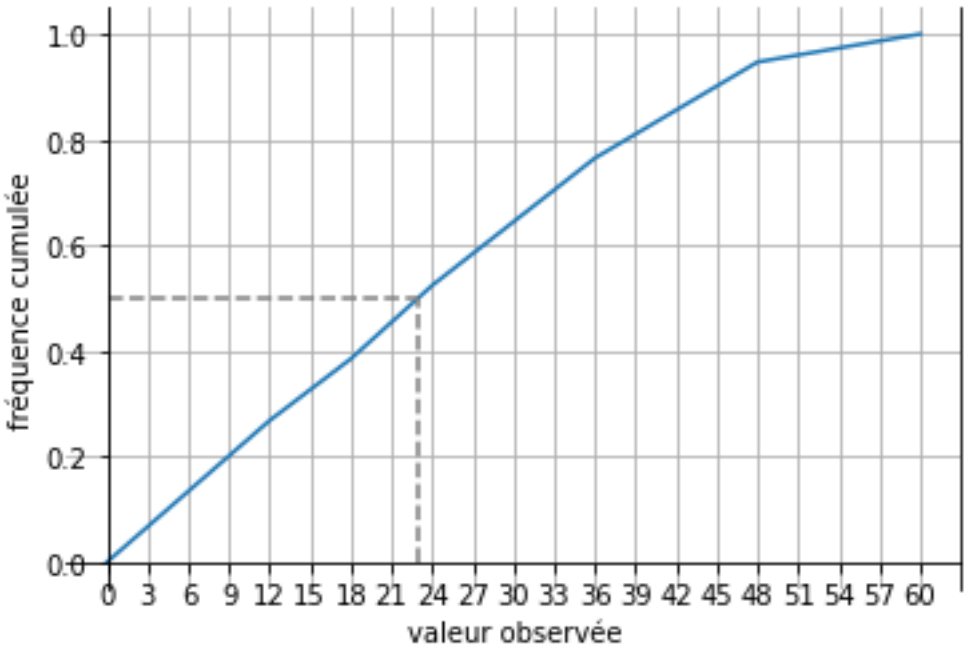
la médiane est alors :

$$m = \frac{a_{i+1} - a_i}{f_i}\left(\frac{1}{2} - \sum_{j=1}^{i-1} f_j\right) + a_i$$

Exemple 26.4 Reprenons le deuxième exemple proposé dans l'exemple 23.1 :

Résultat	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
$[0, 12[$	91	91	0,268	0,268
$[12, 18[$	40	131	0,118	0,385
$[18, 24[$	47	178	0,138	0,524
$[24, 36[$	82	260	0,241	0,765
$[36, 48[$	62	322	0,182	0,947
$[48, 60[$	18	340	0,053	1

Le polygone des fréquences cumulées est le suivant :



Par lecture graphique, on voit que la médiane est proche de 23. Pour trouver la valeur précise, on peut déjà remarquer que la fréquence cumulée dépasse $\frac{1}{2}$ pour la première fois dans l'intervalle $[18, 24[$. On cherche ensuite l'équation de la droite passant par les points de coordonnées $(18, 0.385)$ et $(24, 0.524)$, qui est (en arrondissant à la deuxième décimale) :

$$y = \frac{0.524 - 0.386}{24 - 18}(x - 18) + 0.386 = 0.23\,x - 0.032$$

donc la médiane est $23,13$ car c'est la solution de l'équation

$$0.23\,x - 0.032 = \frac{1}{2}$$

Proposition 26.12

Si x est une série statistique de médiane m et si $y = ax + b$ est une transformation affine de x, alors la moyenne $\overline{y}$ de y vérifie :

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

B.3. Mode

Définition 26.13

i. On appelle mode de la série statistique discrète $((x_i, n_i))_{1 \leqslant i \leqslant p}$ tout réel x_i dont l'effectif n_i est maximal.

ii. On appelle **classe modale** de la série statistique continue $([a_i, a_{i+1}[, n_i)_{1 \leqslant i \leqslant p+1}$ toute classe $[a_i, a_{i+1}[$ dont l'effectif n_i est maximal.

Remarque Une série statistique peut donc avoir plusieurs modes (ou classes modales).

Exemples 26.5

a. Dans le premier exemple de l'exemple 23.1 (lancers de dé), il y a un unique mode, qui est 4 (et dont l'effectif est 15).

b. Dans le deuxième exemple (écrans de téléphone portable), il y a une unique classe modale, qui est $[0, 12[$ (d'effectif 91).

c. Dans le troisième exemple (nombres premiers), il y a deux classes modales, qui sont $[0, 10[$ et $[10, 20[$ (chacune d'effectif 4).

B.4. Quantiles

Définition 26.14

Soit x une série statistique et p un réel appartenant à $[0, 1]$.

1. Si la série est discrète, on appelle **quantile** d'ordre p de x tout réel x tel que la fréquence cumulée des valeurs inférieures ou égales (respectivement supérieures ou égales) à x soit au moins égale à p (respectivement $1 - p$).
2. Si la série est continue, on appelle **quantile** d'ordre p de x l'abscisse du point du polygone des fréquences cumulées dont l'ordonnée est égale à p.

Remarques

a. La médiane est donc un quantile d'ordre $\frac{1}{2}$.

b. Si p est un nombre rationnel de la forme $p = \frac{k}{m}$, un quantile d'ordre p de x est aussi appelé $k^{ème}$ fractile d'ordre m.

Définition 26.15

i. Un quantile d'ordre $\frac{1}{4}$ d'une série statistique est appelé **premier quartile** et noté q_1.

ii. Un quantile d'ordre $\frac{3}{4}$ d'une série statistique est appelé **troisième quartile** et noté q_3.

iii. Un quantile d'ordre $\frac{k}{10}$ ($k \in [\![0, 10]\!]$) est appelé $k^{ème}$ décile.

Remarques

a. Un deuxième quartile est aussi une médiane.

b. En pratique, pour déterminer les quartiles on peut procéder comme pour la médiane. Par exemple, s'il s'agit d'une série statistique discrète dont les valeurs relevées sont, dans l'ordre croissant (au sens large) $x_1, \dots, x_n$, et que l'on cherche à déterminer le premier quartile, son rang hypothétique (s'il fait partie de la série) est $r_1 = \frac{n+3}{4}$ et :

- si $r_1 - \lfloor r_1 \rfloor = 0$ (autrement dit si r_1 est entier), alors il y a un unique premier quartile, égal à x_r,
- sinon, tout réel appartenant à $[x_{\lfloor r \rfloor}, x_{\lfloor r \rfloor + 1}[$ est un premier quartile de x ; comme il est néanmoins souhaitable d'en choisir un (et que tout le monde ne s'accorde pas à choisir le même), donc précisons ici le choix fait par Python :
  - si $r_1 - \lfloor r_1 \rfloor = 0.25$, alors on dit que **le** premier quartile est égal à $\frac{x_{\lfloor r_1 \rfloor} + 3x_{\lfloor r_1 \rfloor + 1}}{4}$,
  - si $r_1 - \lfloor r_1 \rfloor = 0.5$, alors on dit que **le** premier quartile est égal à $\frac{x_{\lfloor r_1 \rfloor} + x_{\lfloor r_1 \rfloor + 1}}{2}$,
  - si $r_1 - \lfloor r_1 \rfloor = 0.75$, alors on dit que **le** premier quartile est égal à $\frac{3x_{\lfloor r_1 \rfloor} + x_{\lfloor r_1 \rfloor + 1}}{4}$,

  On procède de même pour le troisième quartile avec $r_3 = \frac{3n+1}{4}$.

c. Il est plus utile de comprendre que les quartiles permettent de partager la série en quatre séries de même effectif que de chercher à retenir l'indexation précise des termes utilisés dans le calcul des quartiles : ceux-ci seront calculés avec un ordinateur, par exemple avec Python ! On ne s'attardera donc pas à apprendre par cœur une formule inutile.

d. On pourra cependant comprendre la valeur de r si l'on veut qu'un quart des effectifs soient inférieurs ou égaux à x et trois quarts supérieurs ou égaux à x, la valeur de x (si elle fait partie de la série) doit être comptée deux fois, et ceci pour le premier, le deuxième et le troisième quartiles (d'où le « +3 » du rang r quand on partage l'effectif en quatre).

Exemples 26.6

a. Si l'on considère la série statistique $x = (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26)$ (dans ce cas on a donc $n = 13$), la médiane est 14 et les premier et troisième quartile sont 8 et 20 :

$$2 \quad 4 \quad 6 \quad \boxed{8} \quad 10 \quad 12 \quad \boxed{14} \quad 16 \quad 18 \quad \boxed{20} \quad 22 \quad 24 \quad 26$$

On peut constater que $\frac{n+3}{4} = 4$ et que :

- un quart (4 sur 16) des données sont inférieures ou égales à 9.5,
- trois quarts (12 sur 16) des données sont supérieures ou égales à 9.5,

De même $\frac{n+1}{2} = 7$ et :

- la moitié (8 sur 16) des données sont inférieures ou égales à 17,
- la moitié (8 sur 16) des données sont supérieures ou égales à 17,

Enfin $\frac{3n+1}{4} = 12.25$ et :

- un quart (4 sur 16) des données sont supérieures ou égales à 20,
- au moins trois quarts (10 sur 13) des données sont inférieures ou égales à 20.

b. Si l'on considère la série statistique $x = (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32)$ (donc $n = 16$), la médiane est 17 et les premier et troisième quartile sont 9.5 et 24.5 :

$$2 \quad 4 \quad 6 \quad 8 \mid 10 \quad 12 \quad 14 \quad 16 \mid 18 \quad 20 \quad 22 \quad 24 \mid 26 \quad 28 \quad 30 \quad 32$$

On peut constater que $\frac{n+3}{4} = 4.75$ et le premier quartile est égal à 9.5 car :

- au moins un quart (4 sur 16) des données sont inférieures ou égales à 9.5,
- au moins trois quarts (12 sur 16) des données sont supérieures ou égales à 9.5,
- si l'on prend en compte l'intervalle $]8, 10[$, un quart des valeurs sont inférieures à 9.5

De même la médiane est égale à 17 et

- la moitié (8 sur 16) des données sont inférieures ou égales à 17,

- la moitié (8 sur 16) des données sont supérieures ou égales à 17,

Enfin le troisième quartile est égal à 10 car :

- un quart (4 sur 16) des données sont supérieures ou égales à 24.5,
- trois quarts (12 sur 16) des données sont inférieures ou égales à 24.5.

B.5. Déciles

Définition 26.16

Pour tout $i \in [\![0, 10]\!]$, le quantile d'ordre $\frac{i}{10}$ d'une série statistique est appelé $i^{ème}$ décile de la série.

C. Caractéristiques de dispersion

C.1. Étendue et intervalle inter-quartiles

Définition 26.17

Soit x une série statistique.

i. La différence entre la plus grande et la plus petite valeurs de la série est appelée **étendue** de la série.

ii. La différence $q_3 - q_1$ entre le troisième et le premier quartiles est appelée **intervalle inter-quartiles.**

C.2. Variance et écart-type empiriques

Définition 26.18

i. Soit $x = (x_i, n_i)_{1 \leqslant i \leqslant n}$ une série statistique discrète. Pour tout $i \in [\![1, p]\!]$, on note f_i la fréquence de x_i.

- On appelle **variance** de la série x le réel

$$\mathbb{V}(x) = \frac{1}{n} \sum_{i=1}^{p} n_i \left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{p} f_i \left(x_i - \overline{x}\right)^2$$

- On appelle **écart-type** de la série x le réel

$$\sigma(x) = \sqrt{\mathbb{V}(x)}$$

ii. Soit $x = ([a_i, a_{i+1}[, n_i)_{1 \leqslant i \leqslant n}$ une série statistique continue. Pour tout $i \in [\![1, p]\!]$, on note $x_i = \frac{a_i + a_{i+1}}{2}$ le centre de la classe $[a_i, a_{i+1}[$ et f_i sa fréquence.

- On appelle **variance** de la série x le réel

$$\mathbb{V}(x) = \frac{1}{n} \sum_{i=1}^{p} n_i \left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{p} f_i \left(x_i - \overline{x}\right)^2$$

- On appelle **écart-type** de la série x le réel

$$\sigma(x) = \sqrt{\mathbb{V}(x)}$$

Théorème 26.19 ► Formule de Koenig-Huygens

Avec les notations de la définition 26.18, on a :

$$\mathbb{V}(x) = \frac{1}{n} \sum_{i=1}^{n} n_i x_i^2 - \overline{x}^2 = \sum_{i=1}^{n} f_i x_i^2 - \overline{x}^2$$

Exemples 26.7 **a.** Reprenons la série statistique proposée sur les lancers de dés :

Résultat	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
1	11	11	0,183	0,183
2	6	17	0,1	0,283
3	12	29	0,2	0,483
4	15	44	0,25	0,733
5	9	53	0,15	0,883
6	7	60	0,117	1

La moyenne de cette série est :

$$\overline{x} = \frac{11 \times 1 + 6 \times 2 + 12 \times 3 + 15 \times 4 + 9 \times 5 + 7 \times 6}{60} = \frac{103}{30}$$

et la variance est :

$$\begin{aligned}\mathbb{V}(x) &= \frac{11 \times 1^2 + 6 \times 2^2 + 12 \times 3^2 + 15 \times 4^2 + 9 \times 5^2 + 7 \times 6^2}{60} - \left(\frac{103}{30}\right)^2 \\ &= \frac{860}{60} - \left(\frac{103}{30}\right)^2 \\ &= \frac{860 \times 15 - 103^2}{60 \times 15} \\ &= \frac{2291}{900} \\ &\simeq 2,55\end{aligned}$$

b. Considérons la série statistique continue x définie par :

Résultat	Effectif
$[0, 2[$	5
$[2, 4[$	2
$[4, 8[$	4
$[8, 10[$	1

La moyenne de cette série est :

$$\overline{x} = \frac{5 \times 1 + 2 \times 3 + 4 \times 6 + 1 \times 9}{12} = \frac{44}{12} = frac113$$

et sa variance est :

$$\begin{aligned}\mathbb{V}(x) &= \frac{5 \times 1^2 + 2 \times 3^2 + 4 \times 6^2 + 1 \times 9^2}{12} = \frac{44}{12} - \left(\frac{11}{3}\right)^2 \\ &= \frac{248}{12} - \frac{121}{9} \\ &= \frac{62}{3} - \frac{121}{9} \\ &= \frac{62 \times 3 - 121}{9} \\ &= \frac{65}{9} \\ &\simeq 7,2\end{aligned}$$

Proposition 26.20

Si x est une série statistique de variance $\mathbb{V}(x)$ et d'écart-type $\sigma(x)$ et si $y = ax + b$ est une transformation affine de x, alors :

$$\mathbb{V}(y) = a^2\,\mathbb{V}(x) \quad \text{et} \quad \sigma(y) = |a|\,\sigma(x)$$

D. Représentation graphique d'une série statistique univariée

Quand on dispose d'une série statistique x, il est souvent intéressant de résumer la série par un graphique, permettant d'avoir une vision globale de la série au premier coup d'œil. Selon que les données sont regroupées par valeurs ou par classes, on préférera un diagramme en bâtons, une boîte à moustache ou un histogramme.

D.1. Diagrammes en bâtons

Définition 26.21

Soit x une série statistique. Un **diagramme en bâtons** est une représentation graphique dans un , constituée de segments de droite verticaux, d'origine l'axe des abscisses et dont les hauteurs respectives sont proportionnelles aux effectifs (ou aux fréquences) des différentes valeurs (ou classes de valeurs) des caractères.

i. Si la série est une série discrète $(x_i, n_i)_{1\leqslant i\leqslant p}$, on reporte en général les valeurs x_i du caractère en abscisse et leurs effectifs n_i (ou leurs fréquences f_i) en ordonnée.

ii. Si la série est une série continue $([a_i, a_{i+1}[, n_i)_{1\leqslant i\leqslant p}$, on reporte en général les centres $\frac{a_i+a_{i+1}}{2}$ des classes en abscisse et leurs effectifs n_i (ou leurs fréquences) en ordonnée.

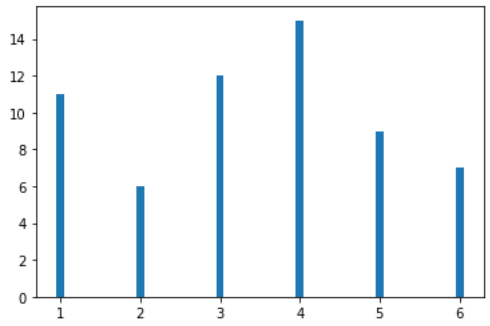
Dans les deux cas, la ligne polygonale obtenue en joignant les sommets des bâtons est appelée **polygone des effectifs** (ou des fréquences) de la série statistique.

Remarques

a. Ce type de diagramme est plus adapté à la représentation de séries discrètes.

b. On peut également représenter une série par un diagramme en bâtons des effectifs cumulés (ou des fréquences cumulées).

Exemple 26.8 Dans le cas de la première série étudiée dans l'exemple 23.1 (lancers de dés), on obtient le diagramme suivant :



D.2. Histogrammes

Définition 26.22

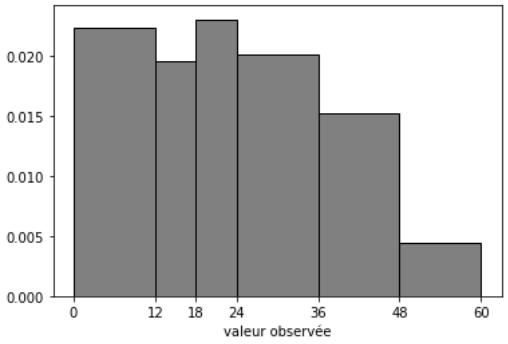
Soit $x = ([a_i, a_{i+1}[, n_i)_{1\leqslant i\leqslant p}$ une série statistique continue. Un **histogramme** est une représentation graphique dans un repère orthogonal, constituée de rectangles dont les bases respectives sont les segments d'extrémités a_i et a_{i+1} sur l'axe des abscisses et dont les **aires** respectives sont proportionnelles aux effectifs (ou aux fréquences) des différentes classes de valeurs des caractères.

Remarque Lorsque les classes ont toutes la même longueur, la hauteur du rectangle de base $[a_i, a_{i+1}]$ peut être égal à n_i mais attention, cela n'est plus vrai si les classes ne sont pas toutes de même longueur car c'est bien l'aire qui est proportionnelle à l'effectif.

Exemples 26.9 **a.** Reprenons la deuxième série de l'exemple 23.1 (les écrans de téléphone portable) :

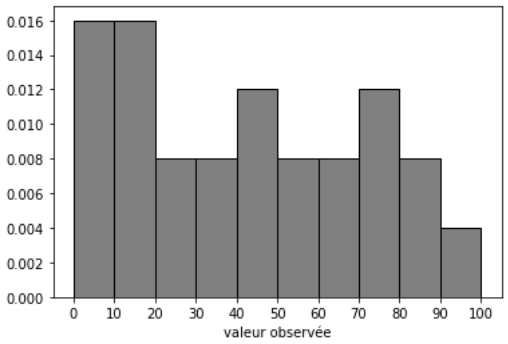
Résultat	Effectif
$[0, 12[$	91
$[12, 18[$	40
$[18, 24[$	47
$[24, 36[$	82
$[36, 48[$	62
$[48, 60[$	18

on obtient le diagramme suivant :



On peut constater dans ce diagramme que bien que la hauteur du premier rectangle est inférieure à celle du troisième, bien que l'effectif de la première classe soit presque le double de celui de la troisième classe : cela s'explique par le fait que la longueur de la première classe est le double de celle de la troisième classe.

b. En reprenant la troisième série de l'exemple 23.1 (les nombres premiers), on obtient le diagramme suivant :



Dans ce cas les hauteurs des différents rectangles sont toutes proportionnelles aux effectifs correspondants, car les classes sont toutes de même longueur (égale à 10).

D.3. Boîtes à moustaches

Définition 26.23

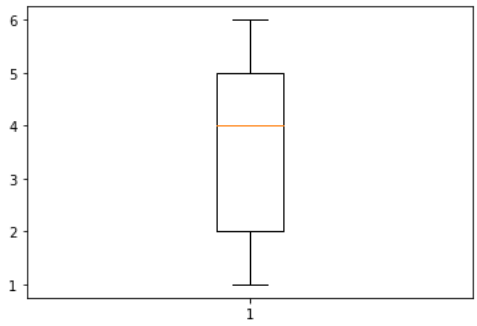
Soit x une série statistique continue. Une **boîte à moustache** (ou diagramme de Tukey) est une représentation graphique résumant la statistique en incluant : le minimum et le maximum, la médiane et les quartiles.

Remarques **a.** Les traits extrêmes représentent le minimum et le maximum.

b. La boîte est délimitée par le premier et le troisième quartile, donc regroupe 50% des valeurs, et elle est coupée à l'intérieur par la médiane.

c. La longueur maximale des traits joignant la boîte aux valeurs extrêmes (appelés « moustaches ») est égale à $1,5$ fois l'intervalle inter-quartiles : lorsque ce n'est pas le cas, les valeurs extérieures sont appelées **valeurs aberrantes**.

Exemple 26.10 Dans le cas de la première série étudiée dans l'exemple 23.1 (lancers de dés), on obtient le diagramme suivant :



Sommaire