Chapitre 2

Paramètres statistiques élémentaires

Les paramètres statistiques concernent principalement les variables quantitatives, et ponctuellement qualitatives. Il existe trois sortes de paramètres : les paramètres de position, les paramètres de dispersion et les paramètres de forme.

2.1 Première approche : espérance, variance et écart type

Pour comprendre les notions paramétriques, on peut les présenter sous la forme d'un exemple simple.

2.1.1 Moyenne

On suppose deux élèves ayant eu les notes suivantes (Tab. 2.1).

Élève	Note 1	Note 2	Note 3	Note 4	Moyenne
1	10	10	10	10	10
2	0	20	5	15	10
Notation statistique	x_1	x_2	x_3	x_4	μ

TABLE 2.1 – Moyenne

La moyenne ne permet pas de percevoir que l'élève n° 1 a été plus régulier que l'élève n° 2. Il faut proposer un autre indicateur pour le mesurer.

2.1.2 Variance

Étape n° 1. Calculer les écarts à la moyenne : $x_i - \mu$.

Élève	$x_1 - \mu$	$x_2 - \mu$	$x_3 - \mu$	$x_4 - \mu$	**Moyenne des écarts**
1	0	0	0	0	0
2	-10	10	-5	5	0

TABLE 2.2 – Écart à la moyenne

Ce n'est pas suffisant. On arrive au même problème que pour la moyenne (Tab. 2.1). Il existe alors deux possibilités : soit prendre la valeur absolue (2a), soit prendre le carré de $x_i - \mu$ (2b)

Étape n° 2a. Calculer la valeur absolue des écarts à la moyenne : $|x_i - \mu|$.

Élève	**Note 1**	**Note 2**	**Note 3**	**Note 4**	**Total**
1	0	0	0	0	0
2	10	10	5	5	30

TABLE 2.3 – Valeur absolue des écarts à la moyenne

Étape n° 2b. Calculer le carré des écarts à la moyenne : $(x_i - \mu)^2$.

Élève	**Note 1**	**Note 2**	**Note 3**	**Note 4**	**Total**
1	0	0	0	0	0
2	100	100	25	25	250

TABLE 2.4 – Carré des écarts à la moyenne

(2b) est appelé la **variance** $\mathbb{V}$. On l'utilise car l'usage du carré offre des propriétés que ne possède pas la valeur absolue.

2.1.3 Généralisation de la variance

Inconsciemment, la moyenne utilisée n'a pas été pondérée. On suppose que toutes les notes ont le même poids p_i

x_i	x_1	x_2	x_3	x_4	**Total**
p_i	0,25	0,25	0,25	0,25	1

TABLE 2.5 – Poids appliqués aux moyennes du tableau n° 2.2

On peut mettre d'autres p_i, mais leur somme doit **toujours** être égale à 1.

Si l'on introduit la notion de variable aléatoire X dont x_1, x_2, x_3 et x_4 sont ses réalisations, le poids correspond à une **probabilité**. Dans le cas de la moyenne simple, on supposait que les notes étaient **équiprobables**.

x_i	x_1	x_2	x_3	x_4
$\Pr(X = x_i)$	p_1	p_2	p_3	p_4

TABLE 2.6 – Notations probabilistes

Étape n° 1. Calculer l'espérance $\mathbb{E}$ de X, $\mathbb{E}(X)$ qui correspond à la moyenne pondérée.

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i \tag{2.1}$$

L'espérance mesure l'ordination de la variable aléatoire X étudiée, c'est-à-dire la **position**. Elle sert de variable de référence.

Étape n° 2. Calculer la variance $\mathbb{V}$ de X, $\mathbb{V}(X)$ qui mesure la moyenne de l'écart à l'espérance au carré.

$$\mathbb{V}(X) = \sum_{i=1}^{n} p_i (x_i - \mathbb{E}(X))^2 \tag{2.2}$$

La variance mesure la régularité de la variable aléatoire X étudiée, c'est-à-dire la **dispersion**. Plus variance est grande, plus X s'écarte de l'espérance, et *vice versa*.

2.1.4 Écart type

L'écart type permet de faire revenir la variance à la même unité que la moyenne ou l'espérance.

$$\sigma(X) = \sqrt{\mathbb{V}(X)} \tag{2.3}$$

De fait, on écrit souvent la variance :

$$(\sigma(X))^2 = \mathbb{V}(X) \tag{2.4}$$

N.B. La racine carrée n'inverse pas l'ordre de la variance, donc son rapport avec la dispersion est le même que la variance. Néanmoins, écart type et variance ne doivent pas être confondus.

Espérance et variance sont des opérateurs. Il existe de fait une algèbre [1] de l'espérance et une algèbre de la variance.

1. Une algèbre est un ensemble de règles de calcul.

2.2 Paramètres de position

Le premier paramètre de position est la moyenne, mais il en existe de nombreux autres : médiane, mode, médiale, *etc*.

2.2.1 La moyenne

Soient $x_1, \ldots, x_p$ les modalités du caractère $n_1, \ldots, n_p$ les effectifs correspondants à p modalités. $n = n_1 + \ldots + n_p$ représente l'effectif total.

Il existe plusieurs manières de calculer les moyennes en fonction de la nature de la variable (Tab. 2.7). On la note : $\bar{x}$ ou encore $\langle x \rangle$.

Nom de la moyenne	**Variables discrètes**	**Variables continues**
Moyenne arithmétique	$\langle x \rangle = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$	$\langle x \rangle = \int_{x_a}^{x_b} x \mathrm{f}\,(x)\,\mathrm{d}x$
	La moyenne ayant été définie par la somme pour une variable discrète, devient une intégrale pour une variable continue.	
Moyenne quadratique	$q = \sqrt[m]{\frac{1}{n}\sum_{i=1}^{n} {x_i}^m}$	
	La moyenne quadratique est peu utilisée, car elle a peu d'intérêt. Exemple d'utilisation. Soit un carré de côté a et soit un autre de côté b, la somme des aires des deux carrés est égale à : $q^2 = \frac{a^2+b^2}{2} \Leftrightarrow q = \sqrt[2]{\frac{a^2+b^2}{2}}$	
Moyenne harmonique	Une condition supplémentaire est nécessaire : $x_i > 0$.	
	$h = \frac{n}{\sum_{i=1}^{p} \frac{n_i}{x_i}}$	$h = \int_{x_a}^{x_b} \frac{1}{x}\mathrm{f}\,(x)\,\mathrm{d}x$
	Exemple d'utilisation. Soit une distance parcourue d dans un sens à la vitesse v_1 et dans l'autre à la vitesse v_2. La vitesse moyenne s'obtiendra en divisant la distance totale $2d$ par le temps mis à la parcourir : $v_i = \frac{2d}{t} \Leftrightarrow t = \frac{2d}{v_i}$ $t = t_1 + t_2 = \frac{d}{v_1} + \frac{d}{v_2}$ $\frac{1}{v} = \frac{1}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right)$	
Moyenne géométrique	Une condition supplémentaire est nécessaire : $x_i > 0$.	
	$g^n = {x_1}^{n_1} \times \ldots \times {x_p}^{n_p} = \prod_{i=1}^{p} {x_i}^{n_i}$	$\ln g = \int_{x_a}^{x_b} (\ln x)\,\mathrm{f}\,(x)\,\mathrm{d}x$
Moyenne mobile (ou glissante)	$MM_s = \frac{M_1+M_2+\ldots+M_n}{n}$	
Moyenne fonctionnelle		$m = \frac{1}{b-a}\int_a^b \mathrm{f}\,(x)\,\mathrm{d}x$

TABLE 2.7 – Les différents types de moyenne

La moyenne[2] arithmétique[3] est sensible aux valeurs extrêmes (ou « aberrantes »). Ainsi, il est parfois nécessaire de les supprimer.

Remarque 1 La moyenne ne change pas si on remplace les effectifs par des effectifs proportionnels.

2. En anglais, on emploie soit *mean* qui est un nom, soit *average* qui est un nom et un adjectif.
3. *arithmetic mean*, *arithmetic average*

Remarque 2 La moyenne ne change pas si on remplace k valeurs $x_1, \ldots, x_k$ affectées des coefficients $n_1, \ldots, n_k$ par leur moyenne partielle affectée de la somme des coefficients $n_1, \ldots, n_k$. Il s'agit d'une **moyenne pondérée** [4]. Soient trois sous-populations ayant pour moyennes partielles $\langle x_1 \rangle$, $\langle x_2 \rangle$ et $\langle x_3 \rangle$, et pour effectifs N_1, N_2 et N_3, alors la moyenne de la population totale est :

$$\langle x \rangle = \frac{N_1 \langle x_1 \rangle + N_2 \langle x_2 \rangle + N_3 \langle x_3 \rangle}{N_1 + N_2 + N_3} \tag{2.5}$$

Il est à noter qu'il existe d'autres moyennes : la moyenne quadratique [5], la moyenne harmonique [6], la moyenne géométrique [7], et la moyenne mobile [8].

2.2.2 La médiane

La **médiane** [9] est la valeur, observée ou possible, dans la série des données classées par ordre croissant (ou décroissant) qui partage cette série en deux parties comprenant exactement le même nombre de données de part et d'autre de cette valeur, notée m_e. On l'appelle également « moyenne du milieu ».

Pour un caractère quantitatif discret On classe les n valeurs de la série statistique par **ordre croissant**.

Deux cas sont possibles.

1. Si n est impair, alors la médiane est le rang de la valeur $\frac{n+1}{2}$.
2. Si n est pair, alors la médiane n'existe pas. On parle d'**intervalle médian** entre les rangs de valeurs $\frac{n}{2}$ et $\frac{n}{2} + 1$. La médiane pratique est souvent le milieu de cet intervalle.

Pour un caractère quantitatif continu La médiane est le nombre m_e tel que la fréquence cumulée jusqu'à ce que la valeur m soit égale à $\frac{1}{2}$.

$$\sum_{i=1}^{m} f_i = \frac{1}{2} \tag{2.6}$$

c'est-à-dire :

4. *weighted average*
5. *root mean square*, *quadratic mean*
6. *harmonic mean*
7. *geometric mean*, *geometric average*
8. *moving average*
9. *median value*

$$\int_{x_a}^{m_e} \mathrm{f}\,(x)\,\mathrm{d}x = \int_{m_e}^{x_b} \mathrm{f}\,(x)\,\mathrm{d}x = \frac{1}{2} \tag{2.7}$$

La médiane est la valeur qui divise la population en deux sous-populations de probabilité équiprobable. De fait, dans la pratique, il s'agit d'une valeur qui ne se calcule pas.

Remarque 1 La médiane n'est pas influencée par les valeurs extrêmes (ou aberrantes), à la différence de la moyenne arithmétique.

Remarque 2 Par contre, la médiane est influencée par le nombre des données. Elle ne peut ainsi être utilisée en théorie de l'estimation. La médiane est déterminée par le **classement** des valeurs, et non par les valeurs extrêmes, donc elle résume bien des distributions fortement dissymétriques.

Remarque 3 La médiane ne peut exister pour une variable statistique discrète, puisqu'elle correspond à la seule valeur possible de cette variable.

Remarque 4 La médiane est le point d'intersection des courbes cumulatives croissantes et décroissantes.

Remarque 5 La médiane ne se prête pas aux combinaisons algébriques. La médiane d'une série globale ne peut pas être déduite des médianes des séries composantes.

2.2.3 Le mode

Le **mode**[10] (ou valeur dominante) m_0 d'une série statistique fait référence à toute modalité correspondant à l'effectif maximal (ou à la densité maximale). Il correspond à la valeur x qui est la plus fréquente (variable discrète), ou qui a la plus forte densité de probabilité :

$$\mathrm{f}'\,(m_0) = 0 \text{ avec } x_a \leq m_0 \leq x_b \tag{2.8}$$

Dit autrement, il s'agit d'une **moyenne de fréquence**.

Remarque 1 Le mode n'existe pas toujours, et, lorsqu'il existe, il n'est pas toujours unique ; on parle alors de distribution bimodale (ou plurimodale).

Remarque 2 Si, après le regroupement des données en classes, deux ou plusieurs modes différents sont trouvés. Il faut considérer que deux ou plusieurs populations distinctes ayant chacune leurs caractéristiques propres sont en présence. Dans ce cas, la moyenne arithmétique n'est pas une caractéristique de tendance centrale.

10. *mode*

2.2.4 La médiale

La **médiale** est la valeur centrale qui partage en deux parties égales la masse de la variable. Il s'agit d'une médiane calculée relativement aux **valeurs globales** $n_i x_i$. Elle est noté m_l. Elle partage les valeurs globales en deux parties égales représentant chacune 50 % des valeurs globales. Le produit $n_i x_i$ ne représente plus seulement l'effectif, mais l'importance de la totalité du caractère possédé par les individus.

Exemple. La médiale partage un ensemble d'employés d'une entreprise en deux groupes, tels que la somme totale des salaires perçus par le premier groupe soit égale à la somme totale des salaires perçus par le second groupe.

Le calcul de la médiale suit trois étapes.

1. Détermination des valeurs globales relatives q_i

$$q_i = \frac{n_i x_i}{\sum_{i=1}^{n} n_i x_i} \tag{2.9}$$

2. Détermination des valeurs globales relatives cumulées croissantes Q_i
3. Détermination de la médiale par interpolation linéaire, c'est-à-dire le calcul classique de la médiane.

La **concentration** C compare la médiale à la médiane m_e.

$$C = \frac{m_l - m_e}{\omega} = \frac{\Delta M}{\omega} \tag{2.10}$$

avec ω l'intervalle de variation de la série statistique (ou étendue).

La médiale est toujours supérieure à la médiane. 50 % des effectifs cumulés croissants ne permettront jamais d'atteindre 50 % de la masse totale.

$$m_l \geq m_e \tag{2.11}$$

La comparaison des valeurs de la médiale et de la médiane constitue une **mesure de concentration**. En général, si ΔM est grand par rapport à l'intervalle de variation, la concentration est forte. Inversement, si ΔM est petit par rapport à l'étendue, la concentration est faible. Dit autrement, lorsque l'écart entre la médiale et la médiane est important par rapport à l'étendue de la distribution de la variable, la concentration est forte. Par contre, lorsque l'écart entre les deux est faible, la concentration est faible ; la distribution est égalitaire. La médiale est liée par l'**indice de C. Gini**.

Corrado Gini (1884-1965)

La courbe de C. Gini a pour objectif de décrire les effets de la concentration d'une population statistique. Elle se construit sur un repère orthonormé à partir des fréquences cumulées relatives. Les valeurs de la fréquence cumulée globale

relative de la série $Q(x)$ sont portées en ordonnée. Elles varient entre 0 et 1. La courbe se construit point par point. On porte sur l'axe des abscisses les valeurs $F(x)$ et sur l'axe des ordonnées les valeurs $Q(x)$. On obtient le **carré de M. O. Lorenz**. Un pourcentage F_i de la population se partage un pourcentage Q_i de la masse totale des valeurs globales. **Plus la courbe s'éloigne de la diagonale du carré, plus la concentration est importante**.

Henrik Antoon Lorentz (1853-1928)

N.B. Les valeurs de la fréquence cumulée de la série (n_i, x_i) sont portées en abscisse. Elles varient entre 0 et 1.

Pour définir n'importe quelle caractéristique, à l'exception de la moyenne arithmétique, il faut que les données soient classées en ordre croissant (ou décroissant). Par exemple, pour le calcul de la médiane, un résultat différent peut être établi selon que les données sont classées par ordre croissant ou décroissant.

2.3 Paramètres de dispersion

Le premier paramètre de dispersion est la variance (ou l'écart type), mais il en existe d'autres : le coefficient de variation, l'étendue, l'écart interquantile, l'écart moyen, la boîte à moustache, *etc*. Les **paramètres de dispersion** correspondent souvent à des paramètres d'échelle concernant les données étudiées. Ils donnent un sens aux paramètres de position.

2.3.1 La variance et l'écart type

Comme vu précédemment, la variance est l'indicateur de dispersion par excellence, mais, exprimé dans la même unité que la moyenne, il est souvent plus pratique d'utiliser l'écart type.

La variance La variance [11] peut être également appelée **écart quadratique moyen** (E.Q.M.) ou variance estimée.

La **variance** $\mathbb{V}$ est la moyenne de la somme des carrés des écarts par rapport à la moyenne arithmétique (ou à la moyenne des carrés moins le carré de la moyenne). Elle correspond au nombre :

11. *variance*

$$\mathbb{V}(X) = \frac{1}{n}\sum_{i=1}^{p}\left[n_i\left(x_i - \langle x\rangle\right)^2\right] = \frac{1}{n}\sum_{i=1}^{p}\left(n_i {x_i}^2\right) - \langle x\rangle^2 \tag{2.12}$$

La variance tient compte de toutes les données ; il s'agit de la **meilleure caractéristique de dispersion**.

L'écart type L'écart type [12] s'appelle aussi **dispersion** ou **déviation standard**.
On note l'écart type $\sigma(X)$ et correspond à la racine carrée de la variance.

$$\sigma(X) = \sqrt{\mathbb{V}(X)} \tag{2.13}$$

Propriété 1 L'écart type caractérise la dispersion d'une série de valeurs. Plus l'écart type est petit, plus les données sont regroupées autour de la moyenne arithmétique, et plus la population est homogène. Cependant, avant de conclure, il faut faire attention à l'ordre de grandeur des données.

Propriété 2 L'écart type permet de trouver le pourcentage de la population appartenant à un intervalle centré sur l'espérance mathématique.

2.3.2 Le coefficient de variation

Le **coefficient de variation** [13] d'une série statistique est le rapport CV :

$$CV = \frac{\sigma}{\langle x\rangle} \tag{2.14}$$

Il s'agit d'un nombre sans dimension permettant de comparer la dispersion de séries statistiques dont les moyennes sont différentes. Ce paramètre est très usité.

Remarque 1 Le coefficient de variation ne dépend pas des unités choisies.

Remarque 2 Il permet d'apprécier la représentativité de la moyenne arithmétique par rapport à l'ensemble des données.

Remarque 3 Il permet d'apprécier l'homogénéité de la distribution. Une valeur du coefficient de variation inférieure à 15 % traduit une bonne homogénéité de la distribution.

Remarque 4 Il permet de comparer deux distributions, même si les données ne sont pas exprimées avec la même unité, ou si les moyennes arithmétiques des deux séries sont différentes.

12. *standard deviation*
13. *coefficient of variation*

2.3.3 L'étendue

L'**étendue** [14] E d'une série statistique associée à un caractère quantitatif est la différence entre la plus grande valeur observée et la plus petite.

$$E = x_{\max} - x_{\min} \tag{2.15}$$

Remarque 1 L'étendue est facile à calculer.

Remarque 2 L'étendue ne contient que des valeurs extrêmes de la série. Elle ne dépend ni du nombre, ni des valeurs intermédiaires. Elle est très peu utilisée dès que le nombre de données dépasse 10.

2.3.4 L'écart interquantile

Les **quantiles** [15] sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.

En partageant la série ordonnée des résultats en quatre parties de même effectif $(k = 4)$, on obtient les quartiles Q_1, Q_2 et Q_3. Le deuxième quartile Q_2 est la médiane. L'écart interquartile est le nombre $Q_3 - Q_1$; il contient 50 % des valeurs de la série.

L'intervalle interquartile peut également se définir de manière continue :

$$\int_{x_a}^{Q_1} \mathrm{f}\,(x)\,\mathrm{d}x = \int_{Q_1}^{x_b} \mathrm{f}\,(x)\,\mathrm{d}x = 0,25 \tag{2.16}$$

L'intervalle interquartiles $(Q_2 - Q_1)$ contient 50 % de la population.

L'**étendue interquartile** [16] (E.I.Q.) est la différence entre deux quartiles. Elle mesure la dispersion de la moitié centrale des observations.

Remarque. On peut définir également les quatre quintiles $(k = 5)$, les neuf déciles $(k = 10)$, les quatre-vingt dix-neuf centiles $(k = 100)$, *etc*.

2.3.5 L'écart moyen

L'**écart moyen** [17] d'un ensemble $X = \{x_1, x_2, \ldots, x_n\}$ est défini par :

$$EM = \frac{1}{n} \sum_{i=1}^{n} |x_i - \langle x \rangle| \tag{2.17}$$

14. *extent*
15. *quantile*
16. *interquartile ranges* (I.Q.R.)
17. *mean spread*, *average spread*

où $\langle x \rangle$ est la moyenne arithmétique des nombres, et $|x_i - \langle x \rangle|$ est la valeur absolue de la différence entre x_i et $\langle x \rangle$.

Si $x_1, x_2, \ldots, x_k$ ont les fréquences respectives $f_1, f_2, \ldots, f_k$ d'apparition, l'écart moyen s'écrit :

$$EM = \frac{1}{n} \sum_{i=1}^{k} n_i \left| x_i - \langle x \rangle \right| \tag{2.18}$$

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

On peut définir aussi l'écart moyen en fonction des écarts absolus à la médiane, ou à tout autre indicateur de tendance centrale. Une propriété importante de la somme $\sum_{i=1}^{k} |x_i - a|$ est qu'elle est minimale pour la médiane. Cela signifie que l'écart moyen par rapport à la médiane est inférieur à tout autre indicateur.

La terminologie d'**écart moyen absolu** serait mieux appropriée que celle d'écart moyen.

L'**écart moyen** e_a peut également s'écrire de manière continue. Il correspond à la moyenne des valeurs absolues des déviations $(x - \langle x \rangle)$.

$$e_a = \int_{x_a}^{x_b} \left| x - \langle x \rangle \right| \mathrm{f}\,(x)\,\mathrm{d}x \tag{2.19}$$

2.3.6 La boîte de dispersion

J. W. Tukey baptisa la boîte de dispersion [18]. La boîte à moustache permet de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quartiles, par exemple.

John Wilder Tukey (1915-2000)

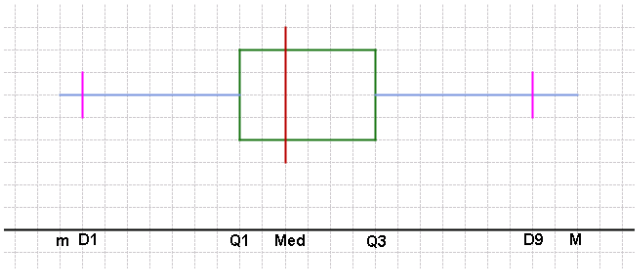
Elle correspond à une représentation graphique d'un **caractère quantitatif**. Elle sert à comparer visuellement plusieurs séries statistiques.

Pour une série donnée, on trace un rectangle qui s'étend de Q_1 à Q_3, et on marque la médiane par un trait. On ajoute les « moustaches » qui sont les segments qui vont de la valeur minimale à Q_1, et de Q_3 à la valeur maximale, puis on fait de même avec les déciles D_1 et D_9 (Fig 2.1).

2.4 Propriétés de l'espérance, de la variance et de l'écart type

Les propriétés suivantes sont universelles. Elles s'appliquent à n'importe quelle distribution. Ces formules permettent d'en établir les paramètres de position et de dispersion. Elles fixent leur algèbre.

18. *box-plot*



Sur la boîte à moustaches ci-dessus, on a représenté le minimum m, le premier décile D1, le premier quartile Q1, la médiane Med, le troisième quartile Q3, le dernier décile D9, et le maximum M.

http://www.bibmath.net/dico/index.php3?action=affiche&quoi=./m/moustache.html

FIGURE 2.1 – La boîte à moustache

2.4.1 Espérance

L'espérance se calcule de façon discrète :

$$\mathbb{E}(X) = \sum_{i=1}^{n} p_i x_i \tag{2.20}$$

ou, en version continue :

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x \tag{2.21}$$

Soient X et Y deux variables aléatoires, soient a et b deux constantes réelles,

Propriété 1. $\mathbb{E}(aX) = a\mathbb{E}(X)$

Propriété 2. $\mathbb{E}(X + b) = \mathbb{E}(X) + b$

Propriété 3. $\mathbb{E}(X \pm Y) = \mathbb{E}(X) \pm \mathbb{E}(Y)$

Propriété 4. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$

Propriété 5. $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$

2.4.2 Variance

La variance se calcule de façon discrète :

$$\mathbb{V}(X) = \sum_{i=1}^{n} p_i \left(x_i - \mathbb{E}(X)\right)^2 \tag{2.22}$$

ou, en version continue :

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 f(x)\,\mathrm{d}x \tag{2.23}$$

La variance peut être vue comme l'espérance du carré des écarts à l'espérance $\mathbb{E}(X)$.

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \tag{2.24}$$

Soient X et Y deux variables aléatoires, soient a et b deux constantes réelles,

Propriété 1 $\mathbb{V}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$

Propriété 2 $\mathbb{V}(aX) = a^2\mathbb{V}(X)$ (fonction scalante)

Propriété 3 $\mathbb{V}(X + a) = \mathbb{V}(X)$

Propriété 4 $\mathbb{V}(aX + b) = \mathbb{V}(aX) = a^2\mathbb{V}(X)$

Propriété 5 Pour $\mathbb{V}(aX) = 0$, si $\mathbb{V}(X) = 0$ alors $X = \mathbb{E}(X)$

2.4.3 Propriétés conditionnées par la nature de la variable aléatoire

Soient X et Y deux variables aléatoires indépendantes, soient a et b deux constantes réelles :

Propriété 1 $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$

Propriété 2 $\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y)$

Propriété 3 $\mathbb{V}(X \pm Y) = \mathbb{V}(X) + \mathbb{V}(Y)$

Attention ! La réciproque est fausse.

Soient X et Y deux variables aléatoires dépendantes, soient a et b deux constantes réelles :

Propriété 1. $\mathbb{V}(X \pm Y) = \mathbb{V}(X) + \mathbb{V}(Y) \pm 2\mathrm{cov}(X, Y)$ avec $\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$

Propriété 2. $\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathrm{cov}(X, Y)$

2.4.4 Écart type

Soit X une variable aléatoire et a et b deux variables réelles :

Propriété 1 $\sigma(X) = \sqrt{\mathbb{V}(X)}$

Propriété 2 $\sigma(X + b) = \sigma(X)$

Propriété 3 $\sigma(aX) = |a|\,\sigma(X)$

Propriété 4 $\sigma(aX + b) = \sigma(aX) = |a|\,\sigma(X)$

2.4.5 Remarque importante

Tout cela se généralise sans problème avec n variables.

2.5 Les paramètres de forme

Les **paramètres de forme** caractérisent l'aplatissement, la symétrie, *etc*. de la loi de distribution statistique de la variable aléatoire étudiée. Avant de définir les paramètres de forme, il faut expliquer rapidement la notion de moments.

2.5.1 Les moments

Pour $r \in \mathbb{N}$ ou $\mathbb{Z}$ et sous réserve de convergence absolue, le **moment** [19] **d'ordre** r est défini par :

$$m_r = \frac{1}{n} \sum_{i=1}^{p} n_i {x_i}^r \tag{2.25}$$

ou encore,

$$m_r = \int_{x_a}^{x_b} x^r \mathrm{f}\,(x)\,\mathrm{d}x \tag{2.26}$$

Le **moment 1** est l'espérance. La formule du moment généralise la notion de moyenne. La variance est la différence entre le moment d'ordre 2 et le carré du moment d'ordre 1.

Pour tout $r' < r$, si m_r existe alors $m_{r'}$ existe.

Le **moment centré** [20] **d'ordre** r est :

$$\mu_r = \frac{1}{n} \sum_{i=1}^{p} n_i \left(x_i - \langle x \rangle\right)^r \tag{2.27}$$

ou encore,

$$\mu_r = \frac{1}{n} \sum_{i=1}^{p} \left(x - \langle x \rangle\right)^r \tag{2.28}$$

ou encore,

$$\mu_r = \int_{x_a}^{x_b} \left(x_i - \langle x \rangle\right)^r \mathrm{f}\,(x)\,\mathrm{d}x \tag{2.29}$$

19. *moment*
20. *central moment*

Le **moment absolu d'ordre** k (ou moment centré d'ordre k) par rapport à un point a, est égal à, sous réserve de l'existence de l'intégrale :

$$\mathbb{E}\left(|X-a|^k\right) = \int |x-a|^k \,\mathrm{f}\,(x)\,\mathrm{d}x \tag{2.30}$$

La propriété essentielle est la **formule de Königs-Huyghens** :

$$\mathbb{E}\left[(X-a)^2\right] = \mathbb{V}\,(X) + \left[\mathbb{E}\,(X) - a\right]^2 \tag{2.31}$$

Johann Samuel König (1712-1757)
Christian Huygens (1629-1695)

Il peut également s'écrire :

$$x_k = \sum_{i=1}^{p} f_i\,(x_i - a)^k \tag{2.32}$$

ou

$$x_k = \frac{1}{n}\sum_{i=1}^{p} n_i\,(x_i - a)^k \tag{2.33}$$

Les moments permettent de caractériser une distribution.

Ordre	Moment	Moment centré
$r=1$	$m_1 = \int_{x_a}^{x_b} x\mathrm{f}\,(x)\,\mathrm{d}x$ Il correspond à la moyenne.	$\mu_1 = 0$
$r=2$	$m_2 = \int_{x_a}^{x_b} x^2\mathrm{f}\,(x)\,\mathrm{d}x$	$\mu_2 = \int_{x_a}^{x_b} (x - \langle x\rangle)^2\,\mathrm{f}\,(x)\,\mathrm{d}x$ Il correspond à la variance. L'écart-type correspond à : $\sigma = \sqrt{\mu_2}$.

TABLE 2.8 – Les moments d'ordre 1 et 2

Démonstration du moment centré d'ordre 1

$$\mu_1 = \int_{x_a}^{x_b} (x - \langle x\rangle)\,\mathrm{f}\,(x)\,\mathrm{d}x = \int_{x_a}^{x_b} x\mathrm{f}\,(x)\,\mathrm{d}x - \int_{x_a}^{x_b} \langle x\rangle\,\mathrm{f}\,(x)\,\mathrm{d}x = m_1 - m_1\int_{x_a}^{x_b} \mathrm{f}\,(x)\,\mathrm{d}x = 0 \tag{2.34}$$

car $\int_{x_a}^{x_b} \mathrm{f}\,(x)\,\mathrm{d}x$ est la loi de probabilité c'est-à-dire que l'intégrale vaut 1.

Les premiers moments centrés sont liés aux moments non centrés (Tab. 2.9).

Moment centré	Moment non centré
$\mu_2 = m_2 - {m_1}^2$	$m_2 = \mu_2 + {m_1}^2$
$\mu_3 = m_3 - 3m_1m_2 + 2{m_1}^3$	$m_3 = \mu_3 - 3m_1\mu_2 + {m_1}^3$
$\mu_4 = m_4 - 4m_1{m_3}^2 + 6{m_1}^2m_2 - 3{m_1}^4$	$m_4 = \mu_4 - 4m_1\mu_3 + 6{m_1}^2\mu_2 + {m_1}^4$

TABLE 2.9 – Lien entre les moments centrés et les moments non centrés

Exemple du moment centré d'ordre 3

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{p} n_i \left(x_i - \langle x \rangle\right)^3 = \frac{1}{n}\sum_{i=1}^{p} n_i \left({x_i}^3 - 2{x_i}^2 \langle x \rangle + x_i \langle x \rangle^2 - {x_i}^2 \langle x \rangle + 2x_i \langle x \rangle^2 - \langle x \rangle^3\right) \tag{2.35}$$

$$\mu_3 = m_3 - 2m_1m_2 + {m_1}^3 - m_1m_2 + 2{m_1}^3 - \frac{1}{n}\sum_{i=1}^{p} n_i \langle x \rangle^3 \text{ où } \frac{1}{n}\sum_{i=1}^{p} n_i = 1 \tag{2.36}$$

$$\mu_3 = m_3 - 3m_1m_2 + {m_1}^3 + 2{m_2}^3 - {m_1}^3 \tag{2.37}$$

$$\mu_3 = m_3 - 3m_1m_2 + 2{m_1}^3 \tag{2.38}$$

2.5.2 Les coefficients β_1 et β_2 de Pearson et de Fisher

Karl Pearson (1857-1936)

Ronald Aymer Fisher (1890-1962)

Pour une **distribution symétrique**, le mode, la moyenne arithmétique et la médiane sont égaux.

La mesure de la dissymétrie β_1 La mesure de la dissymétrie [21] β_1 (ou de l'asymétrie) vaut :

$$\beta_1 = \frac{\mu_3}{{\mu_2}^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3} \tag{2.39}$$

Si $\beta_1 > 0$ alors la distribution est étalée sur la droite. La dissymétrie est dite **positive**.

Si $\beta_1 < 0$ alors la distribution est étalée sur la gauche. La dissymétrie est dite **négative**.

Si $\beta_1 = 0$ alors la distribution est **symétrique**.

21. *shewness*

La mesure d'aplatissement β_2 La mesure d'aplatissement [22] β_2 vaut :

$$\beta_2 = \frac{\mu_4}{{\mu_2}^4} - 3 \tag{2.40}$$

Si $\beta_2 > 0$ alors la distribution est dite **platicurtique**.

Si $\beta_2 < 0$ alors la distribution est dite **leptocurtique**.

Si $\beta_2 = 0$ alors la distribution est dite **mésocurtique**. On peut citer la **loi normale** qui entre dans cette catégorie.

N.B. Si les coefficients d'asymétrie et d'aplatissement sont proches de zéro, alors la distribution est symétrique.

Attention ! Si les coefficients d'asymétrie et d'aplatissement sont des estimateurs biaisés. Il faut utiliser les définitions de la moyenne et de l'écart type non biaisés pour les calculer dans le cadre d'un échantillon.

22. *kurtosis*

Bibliographie

[Béguin, 1979] BÉGUIN, H. (1979). Méthodes d'analyse géographique quantitative. Litec, Paris.

[Chadule, 1997] CHADULE, G. (1997). Initiation aux pratiques statistiques en géographie. Masson, Paris.

[Dumolard, 2011] DUMOLARD, P. (2011). Données géographiques. Analyse statistique multivariée. Lavoisier - Hermès, Paris.

[Dumolard et al., 2003] DUMOLARD, P., DUBUS, N. et CHARLEUX, L. (2003). Les statistiques en géographie. Atout géographie. Belin, Paris.

[Escofier et Pagès, 2016] ESCOFIER, B. et PAGÈS, J. (2016). Analyses factorielles simples et multiples. Cours et études de cas. Sciences sup. Dunod, Paris.

[Fréchet, 1941] FRÉCHET, M. (1941). Sur la loi de répartition de certaines grandeurs géographiques. Journal de la société statistique de Paris, 82:114–122.

[Jacquard, 2000] JACQUARD, A. (2000). Les probabilités. Que sais-je ? n°1571. PUF, Paris. réédition de 1974.

[Korčák, 1940] KORČÁK, J. (1940). Deux types fondamentaux de distribution statistique. Bulletin de l'Institut international de statistique, 30(3):295–299. Rapports et communications présentés à la XXIVe session de l'Institut international de statistique, Prague, 1938, 2e partie.

[Marchand, 1972] MARCHAND, B. (1972). L'usage des statistiques en géographie. L'espace géographique, 1(2):79–100.

[Morgenthaler, 2007] MORGENTHALER, S. (2007). Introduction à la statistique. Enseignement des mathématiques. Presses polytechniques et universitaires romandes, Lausanne. 3e édition augmentée.

[Spiegel, 1984] SPIEGEL, M. R. (1984). Théorie et applications de la statistique. Série Schaum. McGraw-Hill, Paris. réédition de 1972.

[Tenenhaus, 2007] TENENHAUS, M. (2007). Statistique. Méthodes pour décrire, expliquer et prévoir. Dunod, Paris.

[Wonnacott et Wonnacott, 1995] WONNACOTT, T. H. et WONNACOTT, R. J. (1995). Statistique. Économie – Gestion – Sciences – Médecine. Économica, Paris.