Statistique descriptive univariée

Notes de cours

Mohamed Nassiri

Avertissement

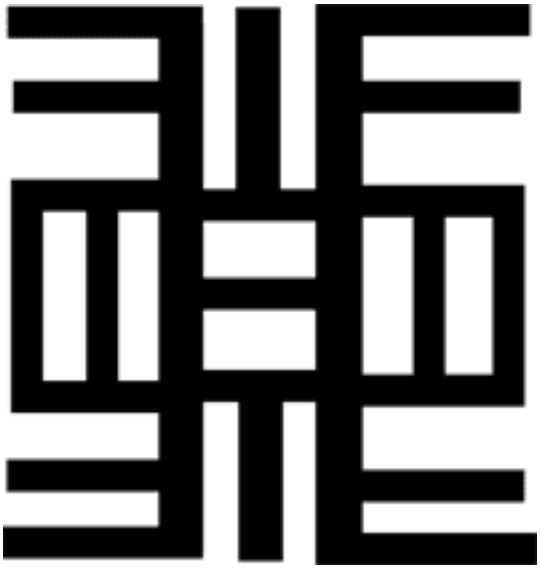
Il s'agit de notes de cours pour l'option mathématiques appliquées aux sciences sociales de Sciences Po Lille (Institut d'Études politiques de Lille).
Le présent document n'a pas la prétention d'être un cours absolument complet.

Copyright © 2022 Mohamed Nassiri

WWW.COQUILLAGESETPOINCARE.FR

Ce document est sous licence *Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International (CC BY-NC-SA 4.0)*.
Voir le Résumé Explicatif : creativecommons.org/licenses/by-nc-sa/4.0/deed.fr
Voir le Code Juridique : creativecommons.org/licenses/by-nc-sa/4.0/legalcode.fr

Dernière version, janvier 2025.



« *Nea onnim no sua a, ohu* [1] »

1. « *Celui qui ne sait pas peut savoir en apprenant* », Adinkra (symbole originaire du Ghana) du savoir, de l'éducation permanente et de la quête continue du savoir

Table des matières

I Présentation de données

1 Présentation de données statistiques 9
1.1 Le tableau à simples entrée 9
1.2 Le diagramme à barres 10
1.3 L'histogramme 11
1.4 Le diagramme circulaire 12
1.5 La boîte à moustache 13
1.6 Le polygone des effectifs ou des fréquences cumulés 14

2 D'autres présentations de données statistiques 17
2.1 La diagramme de Kiviat 17

II Paramètres et indices

3 Les paramètres de position 21
3.1 La moyenne 21
3.2 Le mode 23
3.3 La médiane 23
3.4 Les quartiles et les déciles 25

4 Les paramètres de dispersion 27
4.1 L'étendue 27
4.2 L'écart-interquartile 27
4.3 La variance et l'écart-type 28
4.4 Le coefficient de variation 29

5 Les paramètres de concentration 31
5.1 La courbe de Lorentz 31
5.2 La médiale et l'écart à la médiale 32
5.3 L'indice de Gini 34

6 Les indices simples et synthétiques 37
6.1 Notion de panier 37
6.2 L'indice de Laspeyres 38
6.3 L'indice de Paasche 39
6.4 L'indice de Fisher 39
6.5 Propriétés des indices synthétiques 39
6.6 Tableau récapitulatif 40
6.7 Exemple 40

Présentation de données

1 Présentation de données statistiques . . 9

1.1 Le tableau à simples entrée 9
1.2 Le diagramme à barres 10
1.3 L'histogramme 11
1.4 Le diagramme circulaire 12
1.5 La boîte à moustache 13
1.6 Le polygone des effectifs ou des fréquences cumulés 14

2 D'autres présentations de données statistiques 17

2.1 La diagramme de Kiviat 17

1. Présentation de données statistiques

1.1 Le tableau à simples entrée

Definition 1.1 — Le tableau à simples entrée. Un tableau statistique est un outil utilisé afin de rassembler des données (le plus souvent chiffrées), sous forme de lignes et de colonnes, et de les rendre ainsi plus faciles à utiliser et à interpréter. Le **tableau à simple entrée** possède au moins une colonne, mais une seule ligne.

R On a très souvent des séries statistiques de valeurs x_1, x_2, ..., x_p et d'effectifs respectifs n_1, n_2, ..., n_p que l'on transcrit sous la forme d'un tableau à simple entrée comme ci-dessous :

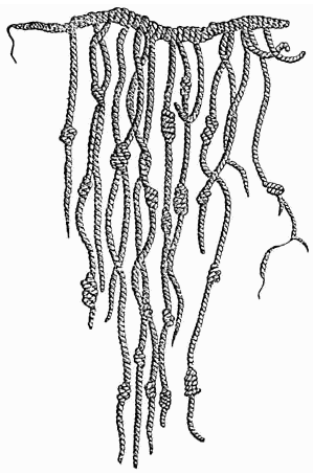
Valeur	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

■ **Exemple 1.1** Le tableau fournit les températures (en °C) moyennes mensuelles à Brest.

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Brest	9,1	9,4	11	12,5	15,6	18,1	20,4	20,6	18,7	15,3	11,9	10

■

R Quipu, quipou, khipu ou quipo, signifie ń noeud ż et ń compte ż en quechua. Le terme désigne aujourd'hui les objets qu'utilisait l'administration inca pour le recensement des données statistiques concernant l'économie et la société de l'empire. En l'absence d'écriture, l'administration figurait les entiers naturels à l'aide de successions de nuds le long de cordelettes de diverses couleurs fixées à une corde : l'ensemble constituait un quipu.
Il est toutefois possible qu'une partie des quipus ait véhiculé une information d'un autre type, notamment des mots-clefs comme payé ou vendu, voire de véritables textes.
Les quipus constituent un système original de consignation de données qui a été développé très tôt dans le Pérou ancien. En effet, certains pourraient dater de quelques milliers d'années comme ceux découverts par Ruth Shady sur le site de la civilisation de Caral remontant à 4 500 ans.

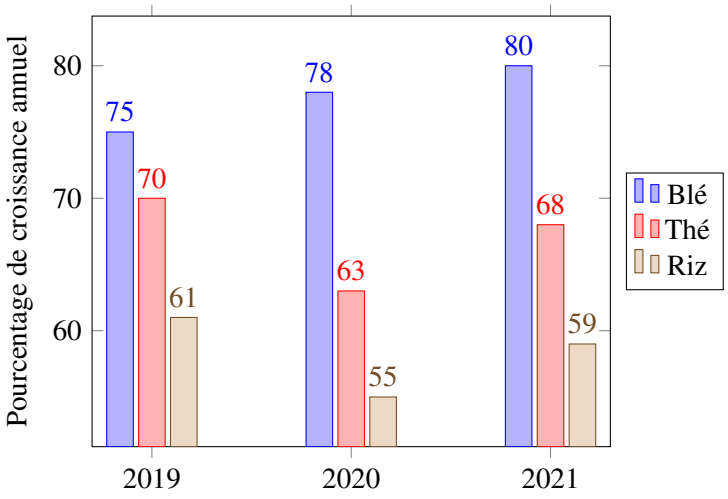


1.2 Le diagramme à barres

Definition 1.2 — Le diagramme à barres. Un **diagramme à barres** (ou en barres), aussi appelé **diagramme à bâtons** (ou en bâtons), est un graphique représentant des variables avec des barres rectangulaires (verticalement ou horizontalement) avec des hauteurs proportionnelles aux valeurs qu'elles représentent.

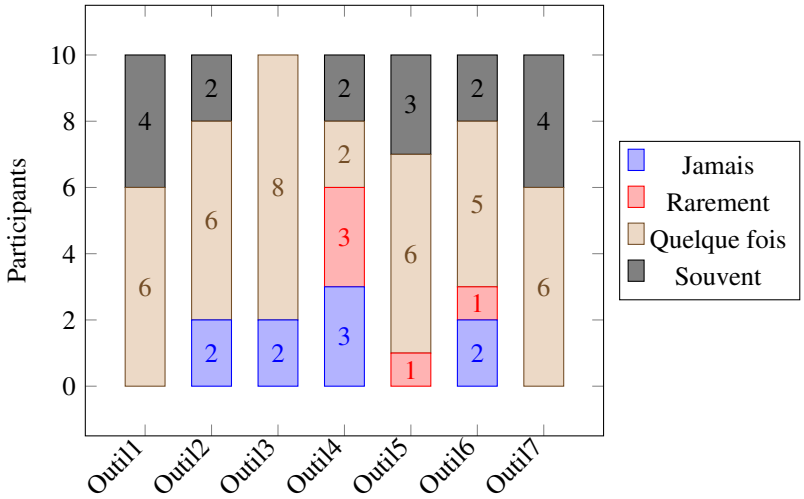
■ **Exemple 1.2** Voici des données statistiques et le diagramme à barres correspondant :

Matières	Blé	Thé	Riz
Quantité en 2019 (en %)	75	70	6
Quantité en 2020 (en %)	70	63	6
Quantité en 2021 (en %)	61	55	59



■

R Il existe une variante que l'on appelle **diagramme à barres empilées**.
Dans l'exemple ci-dessous, on a demandé à 10 personnes leur habitudes de consommation concernant 10 outils numériques.



1.3 L'histogramme

Definition 1.3 — L'histogramme. Un **histogramme** est un moyen de représenter une série statistique dont le caractère est quantitatif continu. Si la série statistique est donnée par les classes ($[a_i, a_{i+1}[$), il est constitué par des rectangles dont la base est le segment $[a_i, a_{i+1}[$ (sur l'axe des réels) et l'aire est proportionnelle à l'effectif de la classe.

R Il faut bien noter que c'est l'aire qui doit être proportionnelle à l'effectif de la classe et non la hauteur elle-même. Si toutes les classes ont la même étendue, cela n'a pas d'importance. Sinon, il faut procéder de la façon suivante : on note ni l'effectif de la classe $[a_i, a_{i+1}[$. On choisit un rapport de proportionnalité k. La hauteur du rectangle de base $[a_i, a_{i+1}[$ sera $\dfrac{k \times n_i}{a_{i+1} - a_i}$.

■ **Exemple 1.3** On a demandé la taille des élèves dans une classe de 33 élèves. On obtient les résultats suivants :

Taille (en cm) :	$150-160$	$160-170$	$170-175$	$175-180$	$180-190$	$190-200$
Effectif :	3	12	9	6	2	1

En choisissant un rapport de proportionnalité $k = 10$.

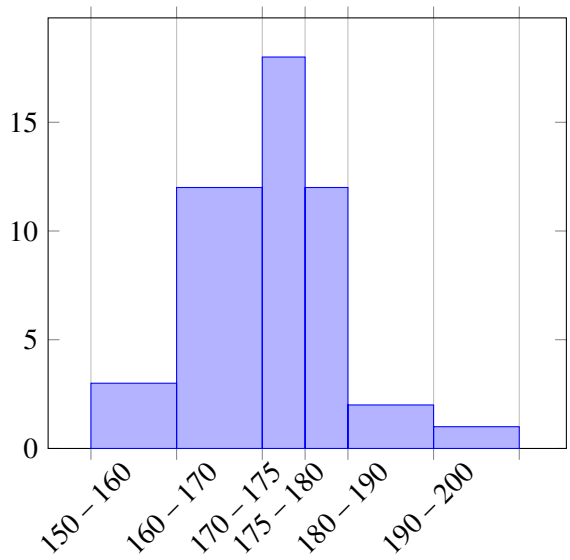
— La hauteur du rectangle de base $[150, 160[$ sera $\dfrac{10 \times 3}{160 - 150} = 3$.

— La hauteur du rectangle de base $[160, 170[$ sera $\dfrac{10 \times 12}{170 - 160} = 12$.

— La hauteur du rectangle de base $[170, 175[$ sera $\dfrac{9 \times 12}{175 - 170} = 18$.

— On réitère l'opération pour tous les intervalles...

L'histogramme correspondant est donc :



■

1.4 Le diagramme circulaire

Definition 1.4 — Le diagramme circulaire. Un **diagramme circulaire** ou **diagramme en secteurs** est un type de graphique utilisé en statistiques. Il permet de représenter un petit nombre de valeurs (ou de classes) par des angles proportionnels à la fréquence (ou l'effectif) de ces valeurs.

▪ **Exemple 1.4** Aude est une agrégative qui prépare sérieusement le concours de l'agrégation de Sciences Économiques et Sociales (SES). Elle décide de lister, par thème, dans un tableau la quantité de notions qu'elle ne maîtrise pas. Voici son tableau :

Matières	Arithmétique	Géométrie	Algèbre	Algorithmique
Nombre de notions	36	72	144	108

Pour mieux se rendre compte de la répartition des notions, elle décide de faire un diagramme circulaire. Elle doit « convertir » ces données en secteurs d'angle.

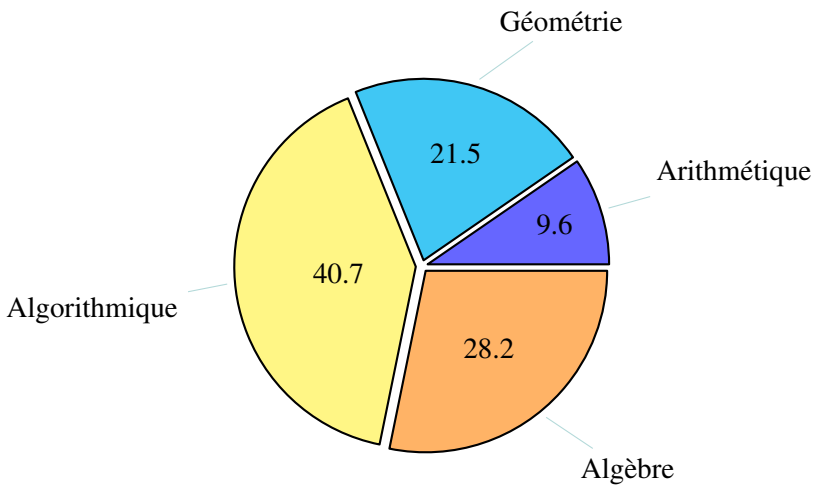
Matières	Arithmétique	Géométrie	Algèbre	Algorithmique	**Total**
Nombre de notions	17	38	72	50	177
Pourcentage	9,6	21,5	40,7	28,2	100
Secteur d'angle	17	38	72	50	177

Pour déterminer les secteurs d'angles, Aude a tout simplement réalisé un produit en croix.
Le nombre total de notions, à savoir 177, correspond à l'angle total, à savoir 360°.

Par exemple, pour l'angle correspondant à la géométrie, on a le calcul suivant :

$$\text{Angle recherché} = \frac{38 \times 360}{177} \simeq 21,5$$

Nombre de notions	**Angle correspondant**
38	Angle recherché
177	360

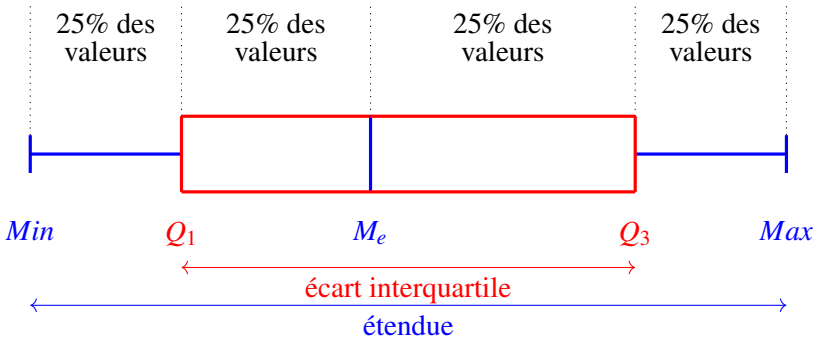


▪

R À cause d'erreurs d'arrondis, il peut arriver que la somme des pourcentages ne donne pas 100, tout comme la somme des angles au centre ne donne pas 360°.

1.5 La boîte à moustache

Definition 1.5 — La boîte à moustache. Il est possible de résumer, sous la forme d'un graphique, l'information fournie par l'étendue, ainsi que par la médiane, les deux quartiles et les intervalles qui les séparent. Ce graphique porte le nom de **boîte à moustaches**, ou encore de **boîte à pattes** ou **diagramme en boîte** (*boxplot* en anglais).



R On parle aussi de **boîte de Tukey** (de son inventeur, John Tukey, en 1977).

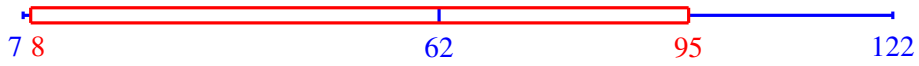
■ **Exemple 1.5** Réalisons la boîte à moustache de la série suivante :

$$73; 122; 73; 8; 95; 43; 64; 44; 32; 111; 115; 7; 7; 8; 111; 60$$

On trouve :

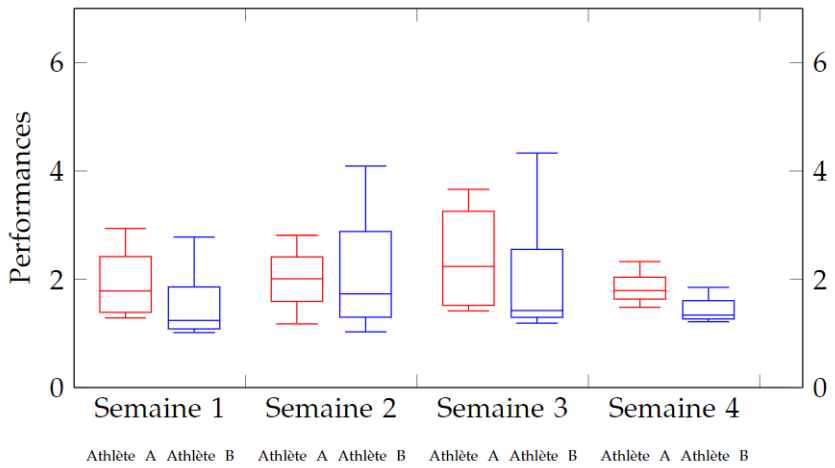
$$Min = 7 \ ; \ Q_1 = 8 \ ; \ Me = 62 \ ; \ Q_3 = 95 \ ; \ Max = 122$$

On a donc



■

R À noter qu'une boîte à moustache seule n'est pas vraiment très intéressante... On les utilise souvent pour comparer plusieurs séries statistiques car c'est un moyen rapide de figurer le profil essentiel de séries statistiques quantitatives.
Par exemple, dans le graphique ci-dessous, nous pouvons comparer rapidement les performances deux athlètes sur 4 semaines.



1.6 Le polygone des effectifs ou des fréquences cumulés

Definition 1.6 L'effectif d'une valeur du caractère étudié est le nombre d'individus de la population ayant cette valeur.
La fréquence d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total de la population. (la fréquence peut être exprimée en pourcentage)

$$\text{fréquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}}$$

On étudie un caractère quantitatif dans une population.

— Les différentes valeurs x_i du caractère quantitatif constituent une série statistique notée (x_i).
— On note n_i l'effectif de la valeur x_i.

Definition 1.7 — Effectif cumulé croissant - Fréquence cumulée croissante.

- **L'effectif cumulé croissant** de la valeur x_i est la somme des effectifs de toutes les valeurs inférieures ou égales à x_i.
- **La fréquence cumulée croissante** de la valeur x_i est la somme des fréquences de toutes les valeurs inférieures ou égales à x_i.

■ **Exemple 1.6** 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	4	9	6	3	1

Le nombre total de journées d'intervention est $2+4+9+6+3+1=25$. Les fréquences des différentes valeurs du nombre d'intervention sont :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Fréquence $f_i = \frac{n_i}{25}$	0,08	0,16	0,36	0,24	0,12	0,04

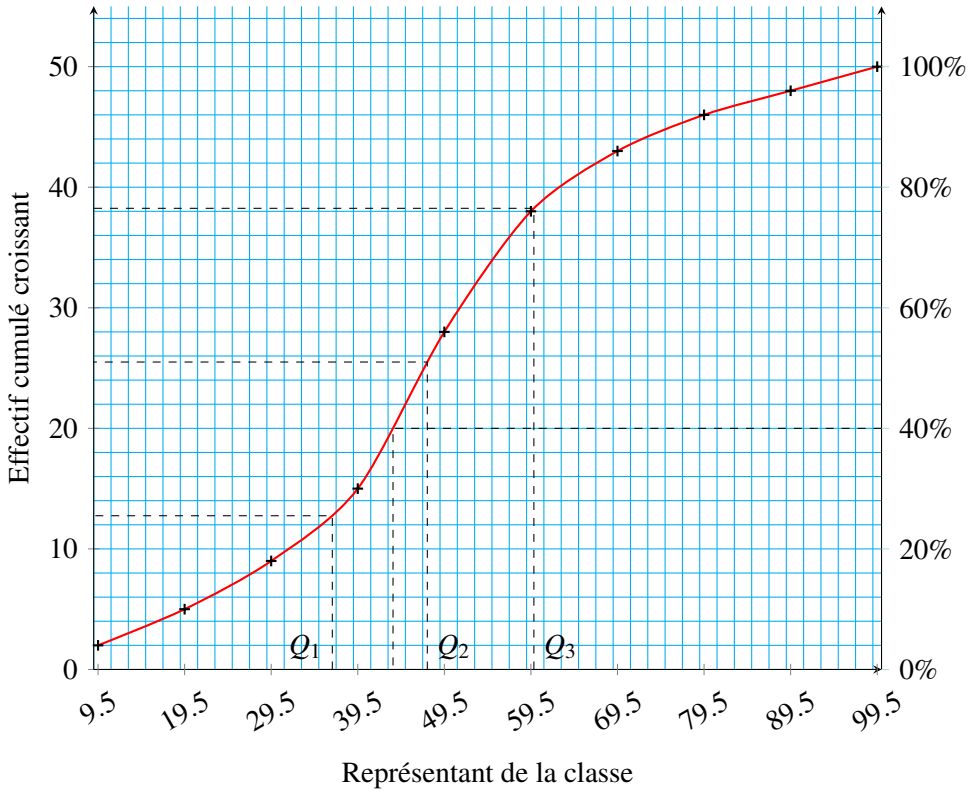
Le tableau suivant donne les effectifs cumulés croissants ainsi que les fréquences cumulées croissantes :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Effectif cumulé	2	6	15	21	24	25
Fréquence cumulée	0,08	0,24	0,6	0,84	0,96	1

■

■ **Exemple 1.7** Réalisons le polygone des effectifs et des fréquences cumulés croissantes

Classe (%)	Effectif	Représentant de la classe	Effectif cumulé croissant
0 – 9	2	9.5	2
10 – 19	3	19.5	5
20 – 29	4	29.5	9
30 – 39	6	39.5	15
40 – 49	13	49.5	28
50 – 59	10	59.5	38
60 – 69	5	69.5	43
70 – 79	3	79.5	46
80 – 89	2	89.5	48
90 – 99	2	99.5	50
Total	50		



■

2. D'autres présentations de données statistiques

2.1 La diagramme de Kiviat

Definition 2.1 — Diagramme de Kiviat. Le **diagramme de Kiviat**, ou **diagramme en toile d'araignée** sert à représenter sur un plan en deux dimensions au moins trois ensembles de données multivariées. Chaque axe, qui part d'un même point, représente une caractéristique quantifiée. Est ainsi facilitée une analyse détaillée de plusieurs objets, ainsi que leur comparaison générale (comparaison des surfaces) ou point par point. Ce type de diagramme n'est utile que si les axes sont correctement normés selon l'importance donnée à chaque caractéristique.

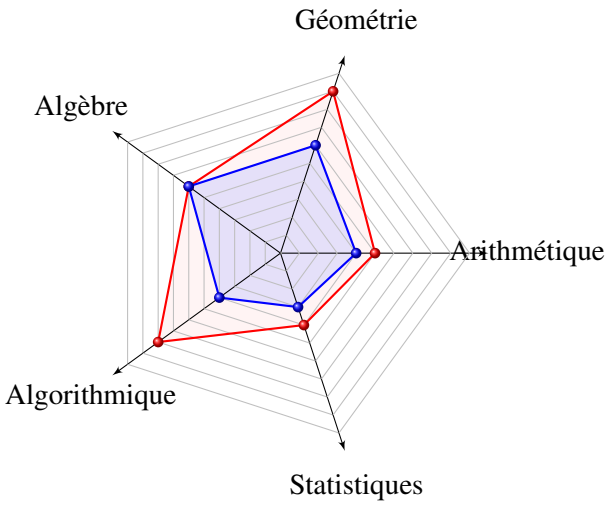
R

— Il est aussi appelé **diagramme en radar** ou encore **diagramme en étoile**.
— Il fut créé en 1877 par le statisticien allemand Georg von Mayr.

■ **Exemple 2.1** Considérons les données suivantes

Matières	Arithmétique	Géométrie	Algèbre	Algorithmique	Statistiques
Notes de l'élève A	5	9	6	8	4
Notes de l'élève B	4	6	6	4	3

Alors le diagramme de Kiviat est



■

II Paramètres et indices

3 Les paramètres de position 21
3.1 La moyenne 21
3.2 Le mode 23
3.3 La médiane 23
3.4 Les quartiles et les déciles 25

4 Les paramètres de dispersion 27
4.1 L'étendue 27
4.2 L'écart-interquartile 27
4.3 La variance et l'écart-type 28
4.4 Le coefficient de variation 29

5 Les paramètres de concentration 31
5.1 La courbe de Lorentz 31
5.2 La médiale et l'écart à la médiale 32
5.3 L'indice de Gini 34

6 Les indices simples et synthétiques . . . 37
6.1 Notion de panier 37
6.2 L'indice de Laspeyres 38
6.3 L'indice de Paasche 39
6.4 L'indice de Fisher 39
6.5 Propriétés des indices synthétiques 39
6.6 Tableau récapitulatif 40
6.7 Exemple 40

3. Les paramètres de position

3.1 La moyenne

Definition 3.1 La **moyenne pondérée** $\overline{x}$ de la série statistique de valeurs x_1, x_2, ..., x_p et d'effectifs respectifs n_1, n_2, ..., n_p est le nombre :

Valeur	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

$$\overline{x} = \frac{n_1 \times x_1 + n_2 \times x_2 + \ldots + n_p \times x_p}{n_1 + n_2 + \ldots + n_p}$$

R

- L'effectif total est $n = n_1 + n_2 + \ldots + n_p$.
- La moyenne est plus sensible aux valeurs extrêmes que la médiane.
- La moyenne peut aussi être notée à l'aide de son initiale m, M ou avec la lettre grecque correspondante μ.

■ **Exemple 3.1** Calcul de la moyenne pondérée de la série suivante

Valeur	6	9	13	15	16
Effectif	2	1	3	2	1

$$\overline{x} = \frac{6 \times 2 + 9 \times 1 + 13 \times 3 + 15 \times 2 + 16 \times 1}{2 + 1 + 3 + 2 + 1} = \frac{106}{9} \simeq 11,8$$

La moyenne des valeurs pondérées par les effectifs est environ de $11,8$. ■

■ **Exemple 3.2** Calcul de la moyenne pondérée de la série suivante.

Valeur	150	155	160	165	170	175	180	185	190
Effectif	25	23	30	50	40	18	10	3	1

La somme des données de la série est :

$$25 \times 150 \text{ cm} + 23 \times 155 \text{ cm} + \ldots + 3 \times 185 \text{ cm} + 190 \text{ cm} = 32860 \text{ cm}$$

L'effectif total de la série est :

$$25 + 23 + 30 + 50 + 40 + 18 + 10 + 3 + 1 = 200$$

Donc la moyenne de la série est égale à :

$$\frac{32860 \text{ cm}}{200} = 164,3 \text{ cm}$$

■

■ **Exemple 3.3** Calcul de la moyenne pondérée de la série suivante. Il s'agit de notes obtenues par une classe de 25 élèves en mathématiques. Les notes sont regroupées en ń intervalles ż ou ń classes ż.

Notes	$[2;8[$	$[8;14[$	$[14;20[$
Effectif	12	8	5

Pour calculer la moyenne générale, on prend la note moyenne de chaque classe (le **centre**) et on l'affecte du nombre d'élèves correspondant :

$$\overline{x} = \frac{5 \times 12 + 11 \times 8 + 17 \times 5}{12 + 8 + 5} = \frac{233}{25} = 9,32$$

La moyenne des notes pondérées par les effectifs est de $9,32$. ■

Proposition 3.1 — Linéarité de la moyenne. Soient a et b deux nombres réels.
Si une série de valeurs $\{x_i\}_{1 \leq i \leq p}$ a pour moyenne $\overline{x}$, alors la série de valeurs $\{ax_i + b\}_{1 \leq i \leq p}$ a pour moyenne $a\overline{x} + b$.

R Il ne s'agit pas d'une proposition particulièrement difficile ; elle est même plutôt naturelle quand on analyse l'exemple précédent.
Cependant pour démontrer cette proposition, il faut surtout comprendre ce que l'on raconte !
On passe de ce tableau :

Valeur	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

à ce tableau

Valeur	$ax_1 + b$	$ax_2 + b$	...	$ax_p + b$
Effectif	n_1	n_2	...	n_p

En gros, on multiplie **toutes** les valeurs par a, puis on ajoute b. Le proposition dit que la moyenne subit la même transformation !

Preuve. On sait que la moyenne de la série $\{x_i\}_{1 \leq i \leq p}$ est :

$$\overline{x} = \frac{n_1 \times x_1 + n_2 \times x_2 + \ldots + n_p \times x_p}{n_1 + n_2 + \ldots + n_p}$$

Calculons maintenant la moyenne de la série de valeurs $\{ax_i + b\}_{1 \leq i \leq p}$ et notons la $\overline{x'}$. On a :

$$\begin{aligned}
\overline{x'} &= \frac{n_1 \times (ax_1 + b) + n_2 \times (ax_2 + b) + \ldots + n_p \times (ax_p + b)}{n_1 + n_2 + \ldots + n_p} \\
&= \frac{n_1 \times ax_1 + n_1 \times b + n_2 \times ax_2 + n_2 \times b + \ldots + n_p \times ax_p + n_p \times b}{n_1 + n_2 + \ldots + n_p} \\
&= \frac{n_1 \times ax_1 + n_2 \times ax_2 + \ldots + n_p \times ax_p + n_1 \times b + n_2 \times b + \ldots + n_p \times b}{n_1 + n_2 + \ldots + n_p} \\
&= \frac{n_1 \times ax_1 + n_2 \times ax_2 + \ldots + n_p \times ax_p}{n_1 + n_2 + \ldots + n_p} + \frac{n_1 \times b + n_2 \times b + \ldots + n_p \times b}{n_1 + n_2 + \ldots + n_p} \\
&= a \underbrace{\frac{n_1 \times x_1 + n_2 \times x_2 + \ldots + n_p \times x_p}{n_1 + n_2 + \ldots + n_p}}_{=\overline{x}} + b \underbrace{\frac{n_1 + n_2 + \ldots + n_p}{n_1 + n_2 + \ldots + n_p}}_{=1} \\
&= a\overline{x} + b
\end{aligned}$$

■

■ **Exemple 3.4**

- Le professeur décide d'ajouter un point à toutes les copies, alors la moyenne sera augmentée de 1 point.
- Le professeur a noté sur 10 et trouve une moyenne de 6 sur 10. Pour ramener les notes sur 20, il multiplie toutes les notes par 2 et la moyenne sera aussi multipliée par 2, on trouve 12 sur 20 de moyenne.

■

Proposition 3.2 — Moyenne et sous-groupes. On étudie un caractère dans une population partagée en deux sous-groupes d'effectifs respectifs n_1 et n_2.
Si la moyenne du caractère est $\overline{x_1}$ dans le premier sous-groupe et $\overline{x_2}$ dans le second, alors la moyenne du caractère dans la population entière est :

$$\overline{x} = \frac{n_1 \times \overline{x_1} + n_2 \times \overline{x_2}}{n_1 + n_2}$$

■ **Exemple 3.5** Deux sous-groupes de sportifs se rencontrent.
Dans le premier sous-groupe, il y a 20 sportifs et le temps moyen pour courir un 100 mètres est de 11 secondes.
Dans le deuxième sous-groupe, il y a 30 sportifs et le temps moyen est de 12 secondes.
- Dans le groupe de 50 sportifs, le temps moyen m pour courir un 100 mètres est :

$$m = \frac{20 \times 11 + 30 \times 12}{50} = 11,6$$

Un troisième sous-groupe de 10 personnes rejoint ces 50 sportifs. Dans ce sous-groupe, le temps moyen est de 13 s.
Dans le groupe des 60 sportifs, le temps moyen m' pour courir un 100 mètres est :

$$m' = \frac{50 \times 11,6 + 10 \times 13}{60} = 11,8$$

■

3.2 Le mode

R

- Le mode est très peu sensible aux données parasites.
- Le mode n'est pas nécessairement unique.

3.3 La médiane

Definition 3.2 La **médiane** M_e d'une série statistique de n valeurs **rangées par ordre croissant** est :
- la valeur centrale si n est impair,
- la moyenne des deux valeurs centrales si n est pair.

R

- On a au moins 50% des valeurs qui sont supérieures ou égales à M_e.
- Dans le cas où l'effectif total est un nombre pair, il est possible que la médiane ne soit pas une valeur de la série.
- Ne pas oublier de ranger les nombres par ordre croissant !

■ **Exemple 3.6** Cherchons la médiane de la série statistique suivante

2 ; 10 ; 5 ; 30 ; 7 ; 50 ; 8 ; 40 ; 9 ; 70 ; 12 ; 200 ; 15 ; 50 ; 10 ; 5 ; 30 ; 7 ; 50 ; 8 ; 40 ; 9 ; 70 ; 12 ; 200 ; 15 ; 50 ; 10 ; 5 ; 30 ; 7 ; 50 ; 8 ; 40 ; 9 ; 70 ; 12 ; 200 ; 15 ; 50 ; 10 ; 5 ; 30 ; 7 ; 50 ; 8 ; 40 ; 9 ; 70 ; 12 ; 200 ; 15 ; 50

On range les données par ordre croissant :

2 ; 5 ; 5 ; 5 ; 5 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 8 ; 9 ; 9 ; 9 ; 9 ; 10 ; 10 ; 10 ; 10 ; 12 ; 12 ; 12 ; 15 ; 15 ; 15 ; 15 ; 30 ; 30 ; 30 ; 30 ; 40 ; 40 ; 40 ; 40 ; 50 ; 50 ; 50 ; 50 ; 50 ; 50 ; 50 ; 50 ; 70 ; 70 ; 70 ; 200 ; 200 ; 200 ; 200 ;
7 012.

L'effectif total de la série est 52. Or, $52 = 26 + 26$. La 26e donnée est 15. La 27e donnée est 15. Donc la médiane de la série est 15. ■

■ **Exemple 3.7** Cherchons la médiane de la série statistique suivante :

Valeur	2	5	7	8	9	10	12	15	20	30
Effectif	1	1	1	1	1	3	1	2	1	3

L'effectif total de la série est 15. Or, $15 = 7 + 1 + 7$. La médiane de la série est la 8e donnée. Donc la médiane de la série est 10. ■

R Attention ! Quand vous listez les valeurs d'une série statistique comme dans l'exemple précédent, il est important de choisir le bon caractère pour séparer les valeurs.
Lorsqu'il n'y a pas d'ambiguïté (par exemple si l'on ne travaille pas avec des nombres décimaux), on peut séparer les valeurs par une virgule.
Cependant, il vaut mieux éviter ce type de séparateur et de prendre la bonne habitude de toujours utiliser le point-virgule...

R Il ne faut pas confondre la médiane et sa cousine la moyenne. Cette dernière est égale à la valeur totale des montants mesurés, divisée par le nombre d'individus. Contrairement à la médiane, la moyenne est très influencée par les valeurs extrêmes et, dans certains cas, moins représentative.
Pour le comprendre, prenons un exemple.

— **Le calcul absurde, les dix souris et l'éléphant.** Prenons un cas absurde. Nous allons faire la moyenne de la masse de 20 souris qui pèsent chacune 30 grammes et de 1 un éléphant qui pèse 5 tonnes soit 5 millions de grammes. Pour calculer cette moyenne nous faisons $\frac{20 \times 30 + 5000000}{21}$, car il y a 21 animaux, soit 238 123 grammes par animaux, soit 238 kilos par animaux ce qui est une aberration pour une souris, même une grosse souris.

  On pourrait faire remarquer que mon exemple est mauvais, car j'ai mélangé des espèces différentes et c'est vrai, mais l'exemple montre que **lorsqu'il y a quelque chose de trop gros ou de trop petit, la moyenne n'a plus de sens, on a donc besoin d'une nouvelle valeur mathématique, la médiane.**

  La médiane de nos souris et de notre éléphant :

  $$30 \ ; \ 30 \ ; \ 30 \ ; \ 30 \ \ldots \ 5\,000\,000$$

  La médiane est de 30 c'est à dire que notre éléphant a été retiré naturellement du paquet.

— **Salaire médian ou salaire moyen ?**

  On comprend donc que dans certains cas la médiane a bien plus de sens que la moyenne et notamment pour le salaire des Français. Si je cherche à calculer le salaire moyen des Français, je prends tous les Français qui travaillent, j'ajoute leur salaire et je divise par le nombre de français qui travaillent. Cela signifie que les quelques milliardaires que nous avons en France, vont faire grossir le salaire moyen comme l'éléphant la masse des souris. Ces gens-là correspondent à des valeurs extrêmes et ne sont donc pas représentatifs. Si en outre, on calcule le salaire médian, on va trier du plus petit salaire

jusqu'au plus gros éliminant d'un côté les très petits salaires, de l'autre les très gros pour conserver une valeur sans les extrêmes.

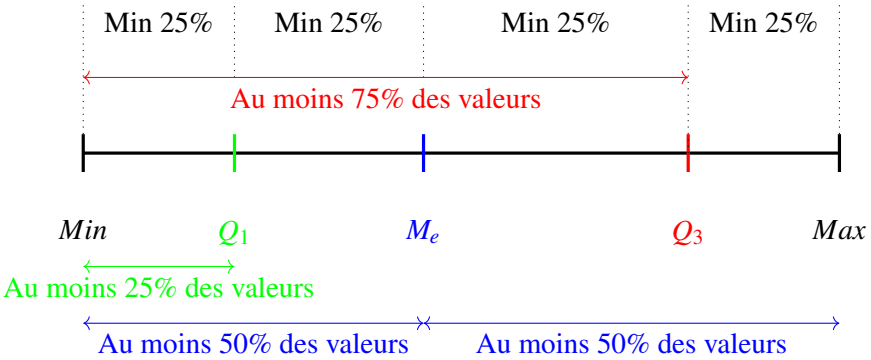
Par conséquent, le salaire moyen est toujours plus élevé que le salaire médian, car les riches rajoutent leur poids dans la balance, si on veut comparer son salaire par rapport au reste des Français, il vaut mieux choisir le salaire médian que moyen.

3.4 Les quartiles et les déciles

Definition 3.3

• Le **premier quartile**, noté Q_1, est la plus petite valeur de la série telle qu'au moins 25% des valeurs soient inférieures ou égales à Q_1.

• Le **troisième quartile**, noté Q_3, est la plus petite valeur de la série telle qu'au moins 75% des valeurs soient inférieures ou égales à Q_3.



R

- Le **rang** du 1er quartile d'une série de n valeurs est le plus entier supérieur ou égal à $\frac{1}{4}n$.
- Le **rang** du 3e quartile d'une série de n valeurs est le plus entier supérieur ou égal à $\frac{3}{4}n$.

■ **Exemple 3.8** On considère la série statistique suivante dont l'effectif total est $n = 10$:

$$2,5;4,8;2;5;5;1;8;9;7;4$$

— D'abord, on range la série dans l'ordre croissant, ce qui donne :

$$1;2;2,5;4;4,8;5;5;7;8;9$$

— Calcul de Q_1 : $\frac{1}{4}n = \frac{1}{4} \times 10 = \frac{10}{4} = 2,5$

On remplace $2,5$ par 3 (on arrondit à l'entier supérieur). Donc le 1er quartile Q_1 est la 3e valeur, c'est-à-dire $2,5$. $Q_1 = 2,5$

— Calcul de Q_3 : $\frac{3}{4}n = \frac{3}{4} \times 10 = \frac{30}{4} = 7,5$

On remplace $7,5$ par 8 (on arrondit à l'entier supérieur). Donc le 3e quartile Q_3 est la 8e valeur, c'est-à-dire 7. $Q_3 = 7$.

■

R Les quartiles font partie d'une famille beaucoup plus grande d'indicateurs de dispersion : **les quantiles**.

En statistiques et en théorie des probabilités, les quantiles sont les valeurs qui divisent un jeu de données en intervalles de même probabilité égale. Il y a donc un quantile de moins que

le nombre de groupes créés. Par exemple, les quartiles sont les trois quantiles qui divisent un ensemble de données en quatre groupes de même probabilité. La médiane quant à elle est le quantile qui sépare le jeu de données en deux groupes de même probabilité.

Les quantiles des multiples du dixième sont des **déciles**. Ils sont d'usage fréquent en géologie minière, en hydrologie, ainsi que dans nombre de statistiques médicales.
Le décile est donc calculé en tant que 10-quantile :

— le seuil du 1er décile sépare les données entre les 10 % inférieurs et le reste des données
— le seuil du 9e décile sépare les 90 % inférieurs des données des 10 % supérieurs.

Par exemple, ce tableau présente le niveau de vie par an en France métropolitaine, en 2007, réparti en déciles. En 2007, le revenu moyen des ménages était de 29 696 €/an.

Tranches de niveau de vie	**Limites des tranches (déciles) en €/an**	**Niveau de vie moyen dans la tranche en €/an**
10 % avaient moins de	10 012 (D1, 1er décile)	7 698 (A)
10 % avaient entre	10 012 et 12 403 (D2)	11 253
10 % avaient entre	12 403 et 14 363 (D3)	13 391
10 % avaient entre	14 363 et 16 201 (D4)	15 297
10 % avaient entre	16 201 et 18 165 (D5)	17 132
10 % avaient entre	18 165 et 20 316 (D6)	19 220
10 % avaient entre	20 316 et 22 969 (D7)	21 565
10 % avaient entre	22 969 et 26 624 (D8)	24 698
10 % avaient entre	26 624 et 33 896 (D9, 9e décile)	29 768
10 % avaient plus de	33 896 (D9)	50 778 (B)
Rapport D9/D1 et B/A	3,39	6,6

— 10 % des Français avaient un niveau de vie inférieur à 10 012 € durant l'année. Ces 10 % de Français avaient, en moyenne, un niveau de vie de 7 698 €. 10 % des Français avaient un revenu compris entre 10 012 € et 12 403 €. Ces Français avaient un niveau de vie moyen de 11 253 €. 10 % des Français avaient un revenu supérieur à 33 896 € (supérieur au 9e décile), et avaient, en moyenne, un niveau de vie de 50 778 €.
— Les 10 % des Français les plus riches avaient, en moyenne, un niveau de vie 6,6 fois supérieur aux 10 % les Français les plus pauvres. Le plancher au-dessus duquel on appartenait aux 10 % les Français les plus riches (le 9e décile, D9) était 3,39 fois supérieur à celui en dessous duquel on appartenait aux 10 % les Français les plus pauvres (le 1er décile, D1).

4. Les paramètres de dispersion

4.1 L'étendue

Definition 4.1 L'**étendue** d'une série statistique est la différence entre la plus grande et la plus petite valeur de la série.

■ **Exemple 4.1** Calcul de l'étendue de la série suivante : $24; 7; 1; 9; 46; 15$.
La plus grande valeur est 46 et la plus petite est 1, donc l'étendue de la série vaut $46 - 1 = 45$. ■

■ **Exemple 4.2**

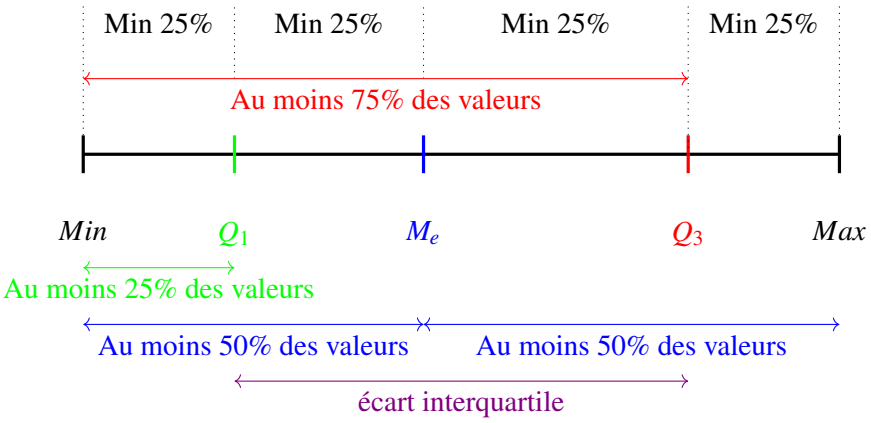
Valeur	2	5	7	8	9	12	15
Effectif	10	30	50	40	70	200	50

L'étendue de la série est égale à $15 - 2 = 13$. ■

R L'étendue ne dépend pas de la taille de la série, mais uniquement des valeurs extrêmes.

4.2 L'écart-interquartile

Definition 4.2
- L'**écart interquartile** est la différence $Q_3 - Q_1$.
- L'**intervalle interquartile** est l'intervalle $[Q_1; Q_3]$.



4.3 La variance et l'écart-type

Definition 4.3 Soit $(x_i;n_i)$, $1 \leqslant i \leqslant p$, une série statistique de moyenne $\bar{x}$ et d'effectif total N.

Valeur	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

- La **variance** de cette série est le nombre V défini par :

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

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

$$\sigma = \sqrt{V}$$

R

- Le calcul de l'écart type se fait généralement à la calculatrice.
- Plus l'écart type est grand, plus les valeurs sont dispersées autour de la moyenne.

■ **Exemple 4.3** On étudie dans une maternité la taille de 50 nouveaux nés.

Taille (en cm)	47	48	49	50	51	52
Effectif	5	8	12	15	9	1

• Calcul de la moyenne $\bar{x}$:

$$\bar{x} = \frac{47 \times 5 + 48 \times 8 + 49 \times 12 + 50 \times 15 + 51 \times 9 + 52 \times 1}{5 + 8 + 12 + 15 + 9 + 1} \simeq 49,36$$

• Disposition pratique de calcul de la variance et de l'écart type (avec la formule de la définition) :

Taille (en cm) x_i	**Effectif** n_i	$(x_1 - \bar{x})^2$	$n_1(x_1 - \bar{x})^2$
47	5	$(47 - 49,36)^2 = 5,5696$	$5 \times 5,5696 = 27,848$
48	8	$(48 - 49,36)^2 = 1,8496$	$8 \times 1,8496 = 14,7968$
49	12	$(49 - 49,36)^2 = 0,1296$	$12 \times 0,1296 = 1,5552$
50	15	$(50 - 49,36)^2 = 0,4096$	$15 \times 0,4096 = 6,144$
51	9	$(51 - 49,36)^2 = 2,6896$	$9 \times 2,6896 = 24,2064$
52	1	$(52 - 49,36)^2 = 6,9696$	$1 \times 6,9696 = 6,9696$
Sommes	50		81,52

La variance est $V = \dfrac{81,52}{50} \simeq 1,6304$ et l'écart type est donc $\sigma = \sqrt{\dfrac{81,52}{50}} \simeq 1,277$ ■

Proposition 4.1 Soit $(x_i;n_i)$, $1 \leqslant i \leqslant p$, une série statistique de moyenne $\bar{x}$ et d'effectif total N. La variance de cette série est le nombre V défini par :

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

Preuve.

$$
\begin{aligned}
V &= \frac{n_1 \times (x_1 - \overline{x})^2 + n_2 \times (x_2 - \overline{x})^2 + \cdots + n_p \times (x_p - \overline{x})^2}{N} \\
&= \frac{n_1 \times \left({x_1}^2 - 2x_1\overline{x} + \overline{x}^2\right) + n_2 \times \left({x_2}^2 - 2x_2\overline{x} + \overline{x}^2\right) + \cdots + n_p \times \left({x_p}^2 - 2x_p\overline{x} + \overline{x}^2\right)}{N} \\
&= \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - 2\frac{n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p}{N} \times \overline{x} + \frac{n_1 + n_2 + \cdots + n_p}{N} \times \overline{x}^2 \\
&= \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - 2\overline{x} \times \overline{x} + \overline{x}^2 \\
&= \frac{n_1 \times {x_1}^2 + n_2 \times {x_2}^2 + \cdots + n_p \times {x_p}^2}{N} - \overline{x}^2
\end{aligned}
$$

Ainsi, la variance d'une série statistique est égale à la différence entre « la moyenne des carrés » et « le carré de la moyenne ».

■

4.4 Le coefficient de variation

Definition 4.4 Le coefficient de variation C_v est le rapport de l'écart-type σ à la moyenne $\overline{x}$ (*i.e.*)

$$c_v = \frac{\sigma}{\overline{x}}$$

R

- Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande. Il est généralement exprimé en pourcentage. Sans unité, il permet la comparaison de distributions de valeurs dont les échelles de mesure ne sont pas comparables.
- Lorsque l'on dispose de valeurs estimées, le CV rapporte l'écart-type de l'estimation à la valeur de cette estimation. Plus la valeur du coefficient de variation est faible, plus l'estimation est précise.

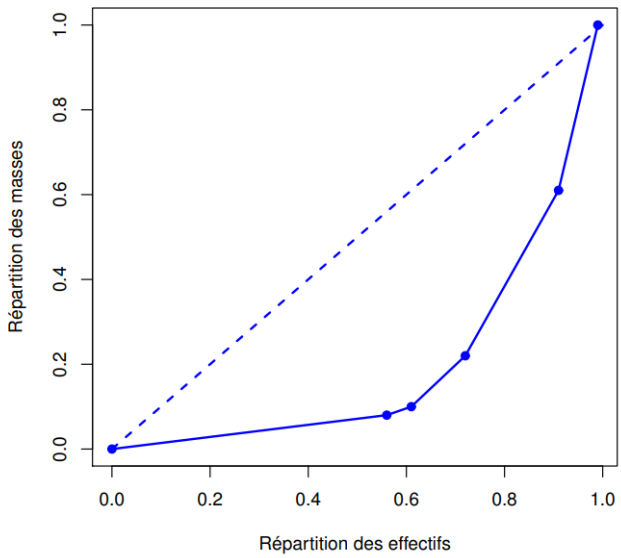
5. Les paramètres de concentration

5.1 La courbe de Lorentz

Elle peut être facilement transposée, notamment à la répartition d'une donnée statistique quelconque, comme :

- les inégalités de répartition d'un actif ou de toute autre distribution de richesse.
- l'état de la répartition des clients au sein d'une clientèle.
- l'état de concentration d'un marché avec la ventilation des parts de marché.

La courbe de Lorenz (économiste américain, 1880-1962) est une représentation graphique qui permet de visualiser graphiquement la répartition des concentrations entre individus et masses.
On calcule les fréquences cumulées des effectifs (qu'on notera p_i) et celles des masses (qu'on notera q_i). On place sur un graphe les points de coordonnées (p_i, q_i) et on les joint par une ligne polygonale.
Cette ligne part du point (0, 0) et se termine au point (1, 1) puisque les fréquences cumulées varient toujours de 0 à 1. Elle est donc inscrite dans le carré de côté 1, parfois appelé le carré de Gini dans ce contexte.



5.2 La médiale et l'écart à la médiale

- Premier exemple d'introduction :
  On peut étudier la répartition des salaires entre les individus à celle des masses salariales qu'ils représentent afin de savoir dans quelle mesure quelques salariés représentent à eux seuls une grande part de la masse salariale (hauts salaires) tandis que beaucoup de salariés représentent une part moindre de la masse totale (bas salaires).
- Deuxième exemple d'introduction :
  On peut étudier la répartition des factures encaissées afin de savoir quelle part de factures représente quelle masse dans le chiffre d'affaire global.
  On suppose que les données sont rassemblées dans un tableau d'effectifs :

Valeurs	v_1	v_2	v_3	$\cdots$	v_k
Effectifs	n_1	n_2	n_3	$\cdots$	n_k

  Les effectifs n_1, n_2, etc. permettent de connaître la répartition des individus en proportions. C'est la distribution des individus.
  Les produits $n_i v_i$ représentent le "poids" de la variable étudiée dans chaque classe C_i, c'est-à-dire pour chaque valeur de v_i. La distribution des $n_i v_i$ est donc la distribution des masses.

On a déjà vu la notion de médiane qui est une valeur telle que 50% de la population soit située en dessous de cette valeur, et 50% au-dessus.

Definition 5.1 — Médiale. La médiale, notée L, est la médiane de la distribution des masses $n_i v_i$.

R Le calcul de la médiale se fait comme celui de la médiane, à partir des fréquences cumulées et nécessite toujours d'effectuer une interpolation. La différence est qu'il s'agit des fréquences cumulées de la distribution des masses et non de celle des individus.

■ **Exemple 5.1** Le tableau suivant donne la répartition des surfaces agricoles utilisées (SAU) dans la région Champagne-Ardennes en 2010 (source INSEE) :

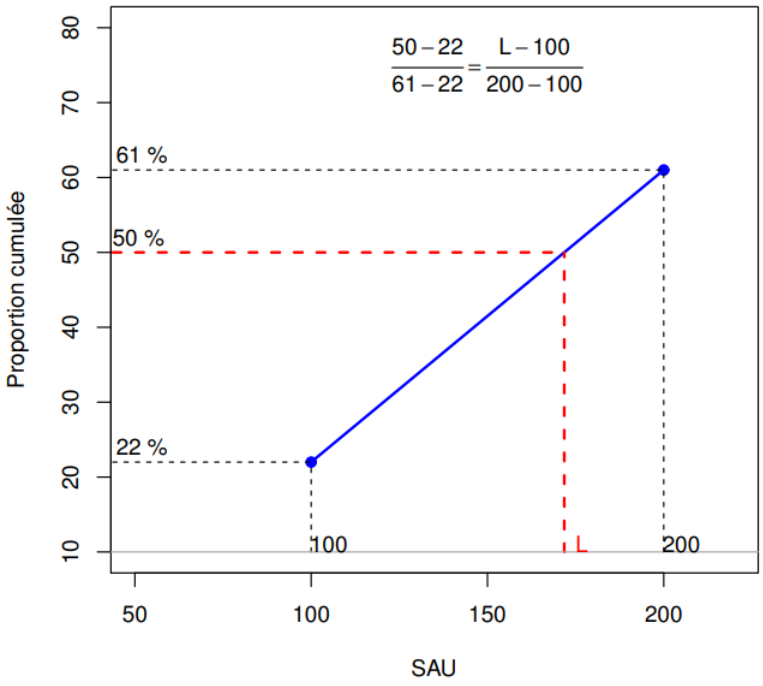
Classes	Effectifs
Moins de 20 ha	13869
De 20 à moins de 50 ha	1224
De 50 à moins de 100 ha	2777
De 100 à moins de 200 ha	4683
De 200 ha à 500 ha	2034

Calculer la médiale.
Chaque classe va être représentée par son milieu v_i. Si n_i est l'effectif de la i-ième classe, on calcule les masses $n_i v_i$ et leur répartition en proportion par rapport à la masse totale $T = \sum n_i v_i$. On en déduit les proportions cumulées. Tous les calculs sont rassemblés dans le tableau suivant.

SAU en ha	Effectif n_i	Milieu v_i	Masse $n_i v_i$	Fréquence $n_i v_i / T$	Fréquence cumulée
$[0,20[$	13869	10	138690	0.08	0.08
$[20,50[$	1224	35	42840	0.02	0.10
$[50,100[$	2777	75	208275	0.12	0.22
$[100,200[$	4683	150	702450	0.39	0.61
$[200,500[$	2034	350	711900	0.39	1.00
Total	**24587**		**1 804155**	1.00	

On cherche où se situe la proportion cumulée de 50%. D'après le tableau, c'est entre $0.22 = 22\%$ et $0.61 = 61\%$. Il faut faire une interpolation linéaire. L'interpolation linéaire consiste à chercher la valeur L qui soit par rapport à 100 et 200 ha comme la valeur 50% par rapport à 22% et 61%.

SAU	Prop. cumulée
100 ha	22%
L?	50%
200 ha	61%



On calcule

$$\frac{50-22}{61-22} = \frac{L-100}{200-100}$$

On en déduit :

$$\frac{28}{39} = \frac{L-100}{100}$$

Finalement :

$$L = 100 + \frac{28 \times 100}{39} = 100 + 71.79 = 171.79\text{ha}$$

La médiale est de 171.79 ha. Cela signifie que toutes les exploitations agricoles dont la SAU est inférieure à cette valeur ont une superficie cumulée qui représente 50% de la superficie totale T. ■

Definition 5.2 — Écart à la médiale. L'écart entre la médiale et la médiane peut donc être considéré comme une mesure de la distorsion de la répartition et sert d'indicateur de concentration. Il est élémentaire à calculer :

$$\Delta M = L - M = \text{ médiale } - \text{ médiane}$$

■ **Exemple 5.2** Dans l'exemple précédent, calculons la médiane. On a besoin de connaître les proportions cumulées des effectifs (et non plus des masses) :

SAU en ha	Effectif n_i	Fréquence n_i/N	Fréquence cumulée
$[0,20[$	13869	0.56	0.56
$[20,50[$	1224	0.05	0.61
$[50,100[$	2777	0.11	0.72
$[100,200[$	4683	0.19	0.91
$[200,500[$	2034	0.09	1.00
Total	**24587**	1.00	

On cherche où se situe la proportion cumulée de 50%. D'après le tableau, c'est entre 0% et $0.56 = 56\%$. On calcule

$$\frac{50-0}{56-0} = \frac{M-0}{20-0}$$

On en déduit :

$$M = \frac{50 \times 20}{56} = 17.86\text{ha}$$

La médiane est de 17.86 ha. Cela signifie que les exploitations agricoles dont la SAU est inférieure à cette valeur constituent 50% de l'effectif total. L'écart entre la médiale et la médiane est :

$$\Delta M = L - M = 171.79 - 17.86 = 153.93$$

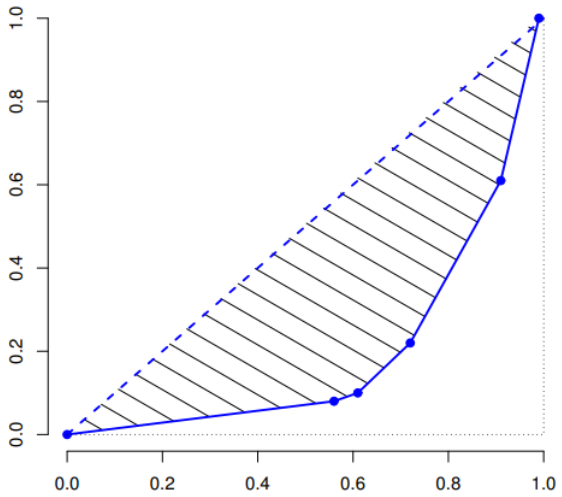
C'est une forte distorsion. On la compare en général à l'étendue des observations qui vaut ici 500 . Le rapport est de $153.93/500 = 0.31 = 31\%$.

■

5.3 L'indice de Gini

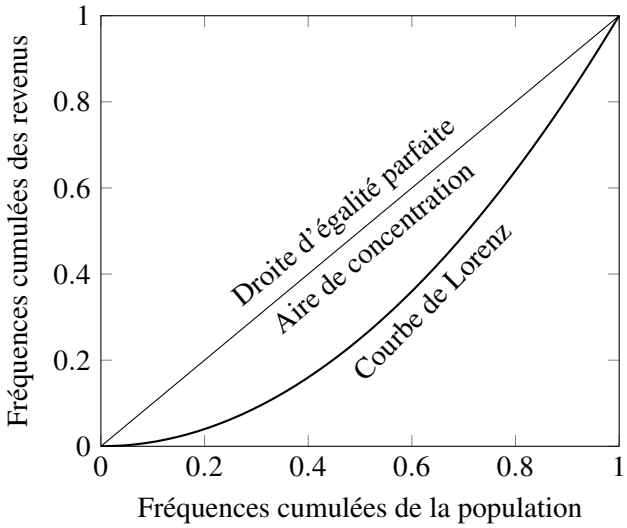
Definition 5.3 — Aire de concentration. L'aire de concentration est la région comprise entre la diagonale et la courbe de Lorenz.

R Plus cette aire est importante, c'est-à-dire plus la courbe de concentration s'écarte de la bissectrice, plus la concentration est forte.



Si on imagine qu'on déforme de plus en plus la courbe de concentration pour l'éloigner de la bissectrice, à la limite elle coïnciderait avec le côté inférieur et le côté droit du carré. L'aire de concentration maximale est donc le triangle inférieur situé sous la bissectrice.

Definition 5.4 — Indice de Gini. L'indice de Gini est le rapport entre la superficie de l'aire de concentration et celle du triangle inférieur du carré



L'indice de Gini permet donc de faire des comparaisons de concentration en les rapportant toutes à la concentration maximale. Cet indice, noté habituellement I_G, est compris entre 0 et 1 :

$$0 \leq I_G \leq 1$$

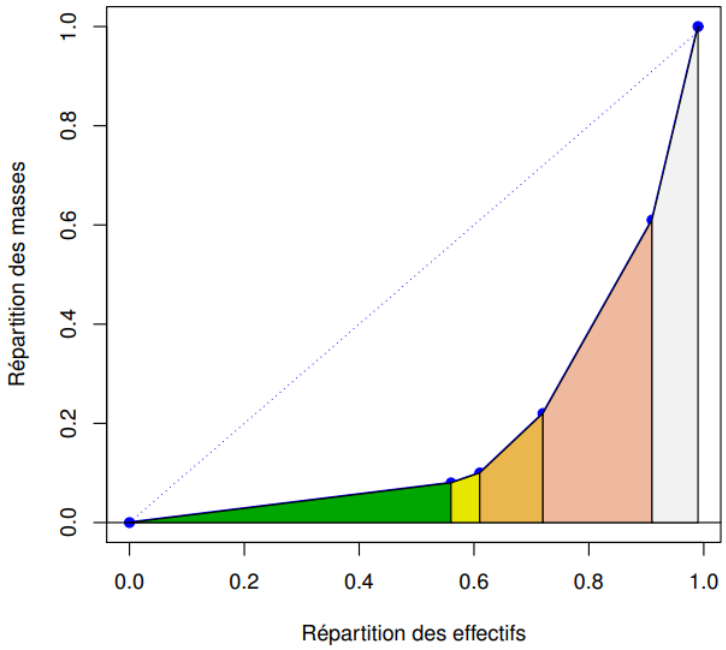
Le carré étant de côté 1 , sa superficie est égale à 1 et donc la superficie du triangle inférieur est égale à $1/2$. Si on note $\mathscr{A}$ l'aire de concentration, on a donc :

$$I_G = \frac{\mathscr{A}}{1/2} = 2 \times \mathscr{A}$$

Maintenant, si on note $\mathscr{B}$ l'aire sous la courbe de concentration, on a $\mathscr{A} = 1/2 - \mathscr{B}$ et par conséquent :

$$I_G = 2 \times (1/2 - \mathscr{B}) = 1 - 2 \times \mathscr{B}$$

La superficie $\mathscr{B}$ est plus facile à calculer car on découpe la région inférieure à la courbe de concentration en trapèzes.



6. Les indices simples et synthétiques

Les indices des prix de Laspeyres-Paasche sont des indices permettant de synthétiser en un indice unique un certain nombre d'indices. Ils représentent le montant nécessaire, dans la période courante, pour acheter un panier de biens et services choisi l'année de référence pour Laspeyres (durant la période courante pour Paasche), rapporté au coût de ce même panier acheté aux prix de l'année de référence.

R Une application courante est le calcul de l'indice des prix à la consommation.

6.1 Notion de panier

Pour définir des indices synthétiques, il faut commencer par choisir un panier de grandeurs qui vont entrer dans la composition de cet indice.
Les indices synthétiques concernent des collections hétérogènes de produits caractérisés par leur prix et leur quantité. On considère donc un panier constitué de n biens :

$$\text{panier } = \left\{\text{Bien}^{(1)}, \text{Bien}^{(2)}, \ldots, \text{Bien}^{(n)}\right\}$$

À la fois, la composition et la valeur de ce panier dépendent du temps. On notera $P_t = \left(p_t^1, p_t^2, \ldots, p_t^n\right)$ le vecteur des prix correspondants et $Q_t = \left(q_t^1, q_t^2, \ldots, q_t^n\right)$ le vecteur des quantités. Produit scalaire
Étant donné deux vecteurs de $\mathbb{R}^n, X = (x_1, x_2, \ldots, x_n)$ et $Y = (y_1, y_2, \ldots, y_n)$, on appelle produit scalaire la quantité

$$x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

Bien qu'on fasse le produit de deux vecteurs, le résultat est un nombre (ou scalaire) : c'est la somme des produits des coordonnées des deux vecteurs. Traditionnellement, le produit scalaire de X et Y est noté " $X.Y$ ". En notation condensée, on peut écrire :

$$X.Y = \sum_{i=1}^{n} x_iy_i$$

Cette notion de produit scalaire va permettre d'écrire la valeur (ou montant) d'un panier.
En effet, si on a n biens de prix $P = (p_1, p_2, \ldots, p_n)$ en quantités $Q = (q_1, q_2, \ldots, q_n)$, la valeur V est

$$V = p_1q_1 + p_2q_2 + \cdots + p_nq_n = \sum_{i=1}^{n} p_iq_i = P.Q$$

On reconnait le produit scalaire des vecteurs P et Q. Autrement dit, on notera simplement dans la suite :

$$V = P.Q$$

Le montant d'un panier est le produit scalaire du vecteur des prix par le vecteur des quantités.
On va maintenant considérer l'évolution du panier dans le temps. On aura donc des prix P_0 et des quantités Q_0 au temps initial, puis des prix P_t et des quantités Q_t au temps t. Il en résulte un montant V_0 au temps inital et un montant V_t au temps t
Une première définition intuitive de l'indice synthétique d'un panier consisterait à prendre l'indice élémentaire de la valeur du panier, autrement dit le rapport

$$I_{t/0} = \frac{V_t}{V_0} \times 100$$

En utilisant la notation des produits scalaires, cet indice s'écrit :

$$\begin{aligned} I_{t/0} &= \frac{P_t \cdot Q_t}{P_0 \cdot Q_0} \times 100 \\ &= \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100 \end{aligned}$$

6.2 L'indice de Laspeyres

Par exemple, pour définir un indice des prix, Étienne Laspeyres (économiste et statisticien allemand, 1834-1913) propose de fixer les quantités à leur valeur initiale et, par conséquent, de définir un contenu du panier qui restera le même en fonction du temps.
Dans ce cas, le vecteur des quantités sera Q_0 aussi bien au départ qu'en tous les temps t qui suivent.
Le montant du panier sera $P_0 \cdot Q_0$ au temps initial. Au temps t, il sera $P_t \cdot Q_0$ (et non pas $P_t Q_t$).
On utilise ici la notation du produit scalaire. Si on veut expliciter les expressions, on a :

$$\begin{cases} P_0 \cdot Q_0 = p_0^1 q_0^1 + p_0^2 q_0^2 + \cdots + p_0^n q_0^n \\ P_t \cdot Q_0 = p_t^1 q_0^1 + p_t^2 q_0^2 + \cdots + p_t^n q_0^n \end{cases}$$

Definition 6.1 — Indice de Laspeyres. L'indice de Laspeyres des prix est donc défini comme le rapport :

$$L_{t/0}^P = \frac{P_t \cdot Q_0}{P_0 \cdot Q_0} \times 100$$

R

- Ici la lettre L est l'initiale de Laspeyres et l'exposant" P signifie prix. L'indice $t/0$ signifie qu'on étudie l'évolution entre le temps 0 et le temps t.
- De la même manière, on peut définir un indice des quantités en fixant le vecteur de prix à sa valeur initiale P_0. On obtient ainsi l'indice de Laspeyres des quantités :

$$L_{t/0}^Q = \frac{P_0 \cdot Q_t}{P_0 \cdot Q_0} \times 100$$

- Ces indices sont exprimés en base 100.
- Si les prix ne varient pas, l'indice de Laspeyres des prix $L_{t/0}^P$ reste égal à 100. Si les quantités ne varient pas, l'indice de Laspeyres des quantités $L_{t/0}^Q$ reste égal à 100.

6.3 L'indice de Paasche

Dans le cas de l'indice des prix, plutôt que de se référer au contenu d'un panier qui a été défini dans le passé (au temps 0), Hermann Paasche (économiste et statisticien allemand, 1851-1925) suggère d'utiliser un contenu défini au temps présent (c'est-à-dire au temps t). L'idée est de considérer que nous sommes au temps t et qu'on retrace l'évolution du panier depuis le temps 0 et calculant le montant qu'il aurait représenté.
Cela signifie que les quantités sont fixées à leur valeur Q_t au temps t.

Definition 6.2 — Indice de Paasche. On obtient ainsi l'indice de Paasche des prix défini comme ceci :

$$P^P_{t/0} = \frac{P_t \cdot Q_t}{P_0 \cdot Q_t} \times 100$$

R De la même manière, on peut définir un indice des quantités en fixant le vecteur de prix à sa valeur au temps t, à savoir P_t et en observant les variations de quantités. On obtient ainsi l'indice de Paasche des quantités :

$$P^Q_{t/0} = \frac{P_t \cdot Q_t}{P_t \cdot Q_0} \times 100$$

R On dit parfois que les indices de Laspeyres sont des indices prospectifs tandis que les indices de Paasche sont des indices rétrospectifs.

6.4 L'indice de Fisher

Aucun des deux types d'indices (Laspeyres ou Paasche) n'est meilleur que l'autre. Le choix de l'un ou de l'autre dépend surtout de la perspective adoptée lorsqu'une étude statistique est effectuée, selon qu'on préfère se référer au temps passé ou au temps présent.
L'indice de Laspeyres a tendance à sur-estimer les tendances tandis que l'indice de Paasche a tendance à sous-estimer. On verra que l'indice de Laspeyres est en général supérieur à celui de Paasche :

$$P \leq L$$

Afin de les réconcilier, Irving Fisher (économiste américain, 1867-1947) a suggéré un autre indice synthétique qu'il considérait comme "idéal".

Definition 6.3 — Indice de Fisher. L'indice de Fisher est la moyenne géométrique des indices de Laspeyres et de Paasche. Il existe un indice de Fisher pour les prix $F^P_{t/0}$ et un autre pour les quantités $F^Q_{t/0}$. Les définitions sont les suivantes :

$$F^P_{t/0} = \sqrt{L^P_{t/0} \times P^P_{t/0}}$$
$$F^Q_{t/0} = \sqrt{L^Q_{t/0} \times P^Q_{t/0}}$$

6.5 Propriétés des indices synthétiques

Proposition 6.1 — Comparaison. Comme il s'agit d'une moyenne, la valeur de l'indice de Fisher est comprise entre le minimum (indice de Paasche) et le maximum (indice de Laspeyres). Autrement dit, on a en général les relations suivantes :

$$P \leq F \leq L$$

Proposition 6.2 — Transférabilité. Aucun des trois indices (Laspeyres, Paasche, Fisher) n'est transférable.

Proposition 6.3 — Réversibilité. Les indices de Laspeyres et de Paasche ne sont pas réversibles.

En fait, quand on renverse les indices de temps, on échange les indices de Laspeyres et de Paasche (à un multiple de 100 près si on est en base 100). On a les relations suivantes :

$$L_{0/t} = \frac{10^4}{P_{t/0}}$$
$$P_{0/t} = \frac{10^4}{L_{t/0}}$$

Ces formules sont valides aussi bien pour les indices de prix que pour les indices de quantités. Les indices de Fisher sont réversibles. On a la relation suivante :

$$F_{0/t} = \frac{10^4}{F_{t/0}}$$

En effet :

$$F_{t/0} = 100 \times \sqrt{\frac{P_t \cdot Q_0}{P_0 \cdot Q_0} \times \frac{P_t \cdot Q_t}{P_0 \cdot Q_t}}$$
$$F_{0/t} = 100 \times \sqrt{\frac{P_0 \cdot Q_t}{P_t \cdot Q_t} \times \frac{P_0 \cdot Q_0}{P_t \cdot Q_0}}$$

On voit qu'en faisant le produit membre à membre de ces deux équations, les termes se simplifient entre les numérateurs et les dénominateurs et il ne reste que :

$$F_{t/0} \times F_{0/t} = 100 \times 100 = 10^4$$

6.6 Tableau récapitulatif

Voici le tableau récapitulatif des différents indices synthétiques :

Indices	Prix	Quantités
Laspeyres	$L^P_{t/0} = 100\frac{P_t \cdot Q_0}{P_0 \cdot Q_0}$	$L^Q_{t/0} = 100\frac{P_0 \cdot Q_t}{P_0 \cdot Q_0}$
Paasche	$P^P_{t/0} = 100\frac{P_t \cdot Q_t}{P_0 \cdot Q_t}$	$P^Q_{t/0} = 100\frac{P_t \cdot Q_t}{P_t \cdot Q_0}$
Fisher	$F^P_{t/0} = \sqrt{L^P_{t/0} \times P^P_{t/0}}$	$F^Q_{t/0} = \sqrt{L^Q_{t/0} \times P^Q_{t/0}}$

6.7 Exemple

Le tableau suivant indique les quantités et les prix unitaires de trois biens constituant un panier de référence :

	Bien 1		Bien 2		Bien 3	
	P	Q	P	Q	P	Q
2000	3.00	5	7.00	10	9.00	7
2005	3.10	5	7.30	8	9.20	8
2010	3.30	6	7.40	5	9.10	10

• **Calculer les indices des prix de Laspeyres et de Paasche pour l'année 2010 en prenant une base 100 en 2000.**

Faisons le calcul détaillé de l'indice de Laspeyres $L^P_{2010/2000}$. Par définition, il faut calculer $100 \times \frac{P_{2010} \cdot Q_{2000}}{P_{2000} \cdot Q_{2000}}$. Le numérateur vaut :

$$P_{2010} \cdot Q_{2000} = (3.30 \times 5 + 7.40 \times 10 + 9.10 \times 7) = 154.2$$

Le dénominateur vaut :

$$P_{2000} \cdot Q_{2000} = (3.00 \times 5 + 7.00 \times 10 + 9.00 \times 7) = 148$$

On en déduit :

$$L^P_{2010/2000} = \frac{100 \times 154.2}{148} = 104.19$$

Faisons maintenant le calcul détaillé de l'indice de Paasche $P^P_{2010/2000}$. Par définition, il faut calculer $100 \times \frac{P_{2010} \cdot Q_{2010}}{P_{2000} \cdot Q_{2010}}$.

Le numérateur vaut :

$$P_{2010} \cdot Q_{2010} = (3.30 \times 6 + 7.40 \times 5 + 9.10 \times 10) = 147.8$$

Le dénominateur vaut :

$$P_{2000} \cdot Q_{2010} = (3.00 \times 6 + 7.00 \times 5 + 9.00 \times 10) = 143$$

On en déduit :

$$P^P_{2010/2000} = \frac{100 \times 147.8}{143} = 103.36$$

Calculons maintenant les indices de quantité. L'indice de Laspeyres des quantités $L^Q_{2010/2000}$ est, par définition, $100 \times \frac{P_{2000} \cdot Q_{2010}}{P_{2000} \cdot Q_{2000}}$.

Le numérateur vaut :

$$P_{2000} \cdot Q_{2010} = (3.00 \times 6 + 7.00 \times 5 + 9.00 \times 10) = 143$$

Le dénominateur vaut :

$$P_{2000} \cdot Q_{2000} = (3.00 \times 5 + 7.00 \times 10 + 9.00 \times 7) = 148$$

On en déduit :

$$L^Q_{2010/2000} = \frac{100 \times 143}{148} = 96.62$$

Faisons maintenant le calcul de l'indice de Paasche pour les quantités. C'est, par définition, $P^Q_{2010/2000} = 100 \times \frac{P_{2010} \cdot Q_{2010}}{P_{2010} \cdot Q_{2000}}$. Le numérateur vaut :

$$P_{2010} \cdot Q_{2010} = (3.30 \times 6 + 7.40 \times 5 + 9.10 \times 10) = 147.8$$

Le dénominateur vaut :

$$P_{2010} \cdot Q_{2000} = (3.30 \times 5 + 7.40 \times 10 + 9.10 \times 7) = 154.2$$

On en déduit :

$$P^Q_{2010/2000} = \frac{100 \times 147.8}{154.2} = 95.85$$

• **Calculer les indices de Fisher en prix et en quantité pour la même période en base 100.**
L'indice de Fisher est la moyenne géométrique des indices de Laspeyres et de Paasche. Pour les prix, on trouve :

$$F^P_{2010/2000} = \sqrt{L^P_{2010/2000} \times P^P_{2010/2000}} = \sqrt{104.19 \times 103.36} = 103.77$$

Pour les quantités, on trouve

$$F^Q_{2010/2000} = \sqrt{L^Q_{2010/2000} \times P^Q_{2010/2000}} = \sqrt{96.62 \times 95.89} = 96.24$$

• **Calculer les indices de prix de Laspeyres $L^P_{2010/2005}$ et $L^P_{2005/2000}$. A-t-on la propriété de transférabilité ?**
On procède comme dans les questions précédentes. Tout calcul fait, on obtient :

$$\begin{cases} L^P_{2010/2005} = 100.68 \\ L^P_{2005/2000} = 103.31 \end{cases}$$

On constate que

$$L^P_{2010/2000} \neq L^P_{2010/2005} \times L^P_{2005/2000}$$
$$\text{car } 100.68 \times 103.31/100 = 104.01 \neq 104.19$$

Les indices de Lapeyres ne sont donc pas transférables. Les indices de Paasche ne le sont pas non plus. • **Même question avec les indices de prix de Fisher $F^P_{2010/2005}$ et $F^P_{2005/2000}$.**
Il faut d'abord calculer les indices de Paasche pour ces deux périodes. On obtient :

$$\begin{cases} P^P_{2010/2005} = 100.48 \\ P^P_{2005/2000} = 103.15 \end{cases}$$

En combinant avec la question précédente, on en tire les indices de Fisher :

$$\begin{cases} F^P_{2010/2005} = \sqrt{L^P_{2010/2005} \times P^P_{2010/2005}} = \sqrt{100.68 \times 100.48} = 100.58 \\ F^P_{2005/2000} = \sqrt{L^P_{2005/2000} \times P^P_{2005/2000}} = \sqrt{103.31 \times 103.15} = 103.23 \end{cases}$$

On avait trouvé $F^P_{2010/2000} = 103.77$ et on constate que

$$100.58 \times 103.23/100 = 103.83 \neq 103.77$$

Les indices de Fisher ne sont donc pas transférables.