Caractéristiques de forme et de concentration

Cette leçon a 2 objectifs essentiels :
1) préciser l'allure de la courbe des fréquences sans avoir recours à son tracé
2) introduire la notion de concentration et montrer comment la déterminer

1. Les caractéristiques de forme

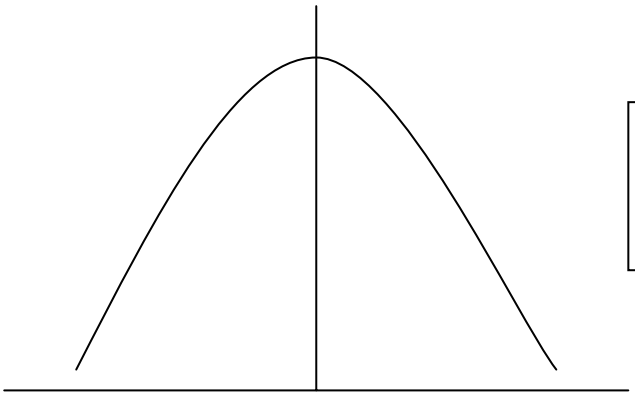
Les caractéristiques de forme permettent de donner l'allure de la courbe des fréquences en ayant recours aux mesures de l'asymétrie et de l'aplatissement .

1.1 Les mesures de l'asymétrie

L'asymétrie d'une distribution s'étudie par rapport à une valeur centrale : il s'agit d'évaluer si la distribution est plus étalée à gauche ou à droite, de cette valeur, ou bien si les observations sont également réparties de part et d'autre de cette valeur.

1.1.1 Distribution symétrique

Une distribution est symétrique si les observations repérées par leurs fréquences sont également dispersées de part et d'autre d'une valeur centrale qui peut être le mode, la médiane ou la moyenne arithmétique.
La moyenne arithmétique, la médiane et le mode se situent directement sous le sommet du polygone de fréquences



Distribution symétrique

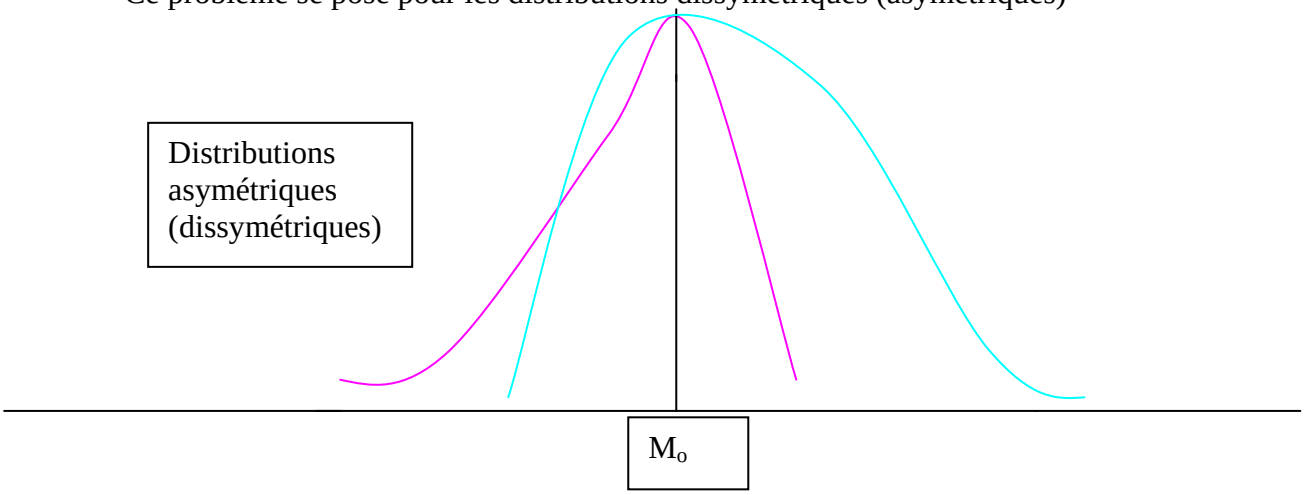
$M_o = M_e = \overline{x}$

Dans une distribution symétrique, les 3 caractéristiques de valeur centrale sont donc confondues :

$$\overline{x} = M_e = M_o$$

Dès lors le problème du choix de la mesure de tendance centrale la plus appropriée pour décrire une distribution ne se pose plus.

Ce problème se pose pour les distributions dissymétriques (asymétriques)



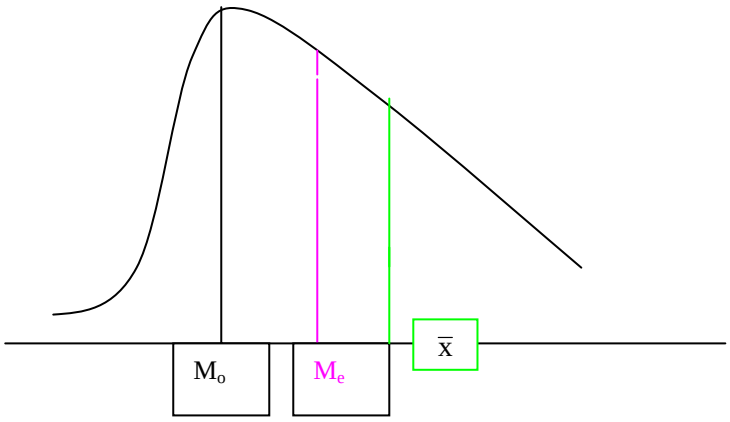
1.1.2 Distribution dissymétrique

Lorsque les distributions deviennent dissymétriques, les mesures de tendance centrale prennent des valeurs différentes : ainsi le mode demeure sous le sommet de la courbe tandis que la moyenne arithmétique se déplace dans le sens de l'étalement (à droite ou à gauche), et la médiane se retrouve entre le mode et la moyenne.
On distingue ainsi 2 formes d'asymétrie :

1.1.2.1 Courbe oblique à gauche ou étalée vers la droite

Une courbe non symétrique est oblique à gauche (étalée vers la droite) si

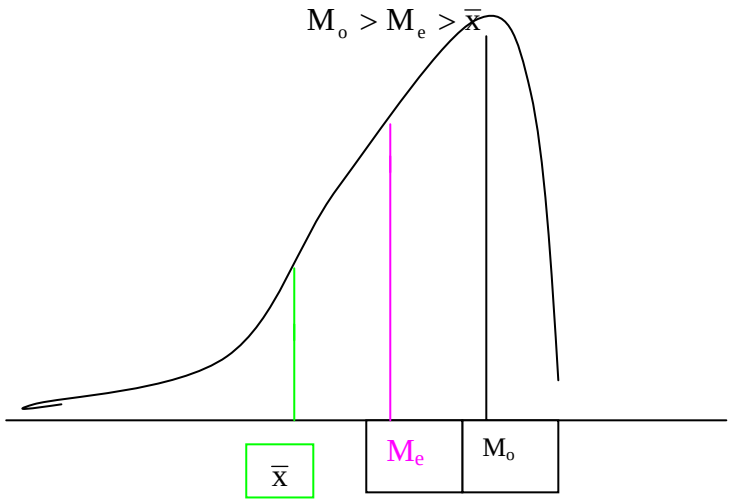
$$\bar{x} > M_e > M_o$$



Ainsi, dans une distribution oblique à gauche (ou étalée vers la droite) (ou encore positivement dissymétrique),
le mode se trouve directement sous le sommet de la courbe et possède la plus petite valeur ;
la moyenne fortement influencée par les valeurs extrêmes se déplace vers ces valeurs et prend la plus grande valeur,
la valeur de la médiane se situe entre les valeurs du mode et de la moyenne.

1.1.2.2 Courbe oblique à droite ou étalée vers la gauche

Une courbe non symétrique est oblique à droite (étalée vers la gauche) si



Ainsi, dans une distribution oblique à droite (ou étalée vers la gauche) (ou encore négativement dissymétrique),

le mode se trouve encore sous le sommet de la courbe et a la plus grande valeur des 3 mesures;
la moyenne fortement influencée a la plus petite valeur,
la valeur de la médiane, comme toujours, se situe entre les valeurs du mode et de la moyenne.

Dans le choix de la mesure de tendance centrale appropriée, nous devons tenir compte des caractéristiques de chacune de ces mesures et du type de données disponibles.

Pour mesurer l'asymétrie, il est possible d'utiliser des coefficients c'est-à-dire des nombres sans dimension.

1.1.3 Les coefficients de l'asymétrie

Il est important de noter que ces coefficients ne peuvent être calculés que si :

1- la distribution ne présente pas plusieurs modes
2- la distribution contient un nombre assez élevé d'observations.

Ces coefficients sont des nombre purs, indépendants des unités de mesure.

1.1.3.1 Coefficient de Yule

Ce coefficient propose une mesure de l'asymétrie en comparant l'étalement vers la gauche et l'étalement vers la droite.
L'étalement vers la gauche et l'étalement vers la droite sont tous deux repérés par la position des quartiles.

$$S=\frac{(Q_3-Q_2)-(Q_2-Q_1)}{(Q_3-Q_2)+(Q_2-Q_1)}$$

Si $S = 0$ la distribution est symétrique ($M_o=M_e=\bar{x}$)
$S > 0$ la distribution est oblique à gauche ou étalée vers la droite ($M_o<M_e<\bar{x}$)
$S < 0$ la distribution est oblique à droite ou étalée vers la gauche ($M_o>M_e>\bar{x}$)

1.1.3.2 Coefficient de Pearson (1)

Le premier coefficient de Pearson n'est valable que pour des distributions faiblement asymétriques. Il analyse la position des 2 valeurs centrales que sont la moyenne arithmétique et le mode, en relativisant cette position par la dispersion de la distribution.

$$s = \frac{\bar{x} - M_o}{\sigma_x}$$

Si $S = 0$ la distribution est symétrique ($M_o=M_e=\bar{x}$)
$S > 0$ la distribution est oblique à gauche ou étalée vers la droite ($M_o<M_e<\bar{x}$)
$S < 0$ la distribution est oblique à droite ou étalée vers la gauche ($M_o>M_e>\bar{x}$)

1.1.3.3 Coefficient de Pearson (2)

Le second coefficient de Pearson s'appuie sur le calcul des moments centrés d'ordre impair.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\mu_3^2}{\sigma_x^6}$$

Si $\beta_1 = 0$ la distribution est symétrique ($M_o=M_e=\bar{x}$)
Si $\beta_1 > 0$ la distribution est oblique à droite ou oblique à gauche selon le signe de μ_3

1.1.3.4 Coefficient de Fisher

Il s'agit tout simplement de la racine carrée du second coefficient de Pearson β_1

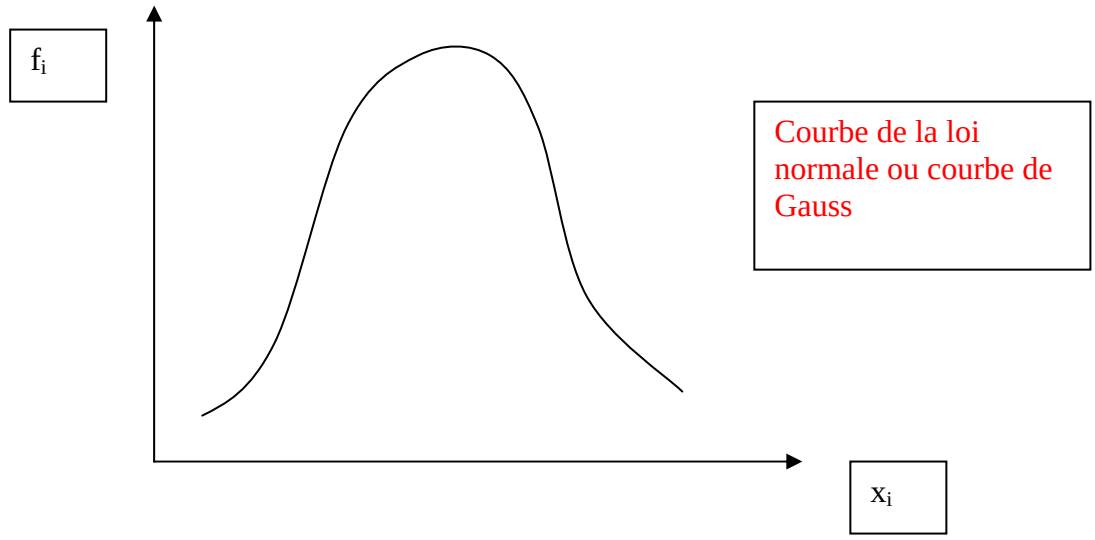
$$\gamma_1 = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \sqrt{\frac{\mu_3^2}{\sigma_x^6}} = \frac{\mu_3}{\sigma_x^3}$$

Si $\gamma_1 = 0$ la distribution est symétrique ($M_o=M_e=\bar{x}$)
$\gamma_1 > 0$ la distribution est oblique à gauche ou étalée vers la droite ($M_o<M_e<\bar{x}$)
$\gamma_1 < 0$ la distribution est oblique à droite ou étalée vers la gauche ($M_o>M_e>\bar{x}$)

1.2 Les mesures de l'aplatissement

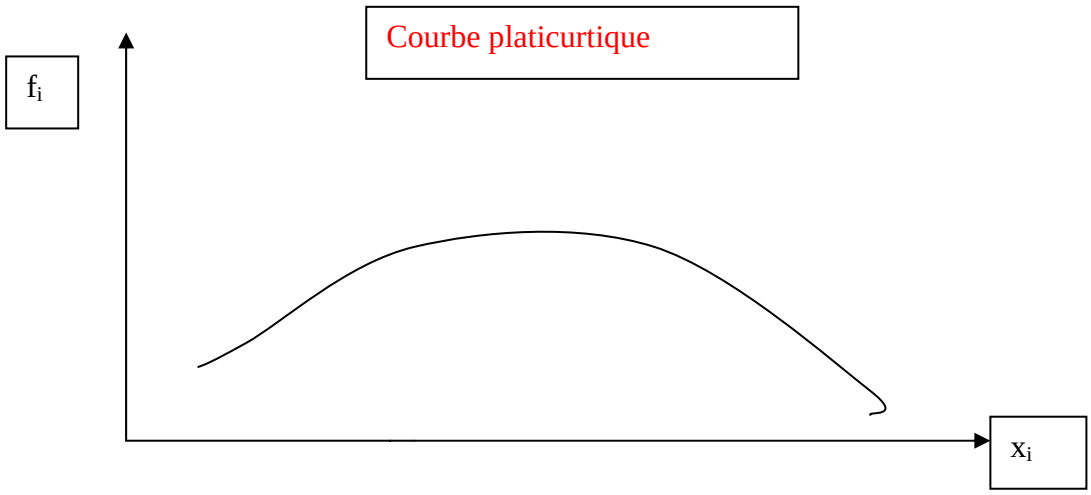
Une distribution peut être plus ou moins aplatie selon qu'une proportion plus ou moins grande des observations est proche de son mode. Lorsqu'une forte proportion des observations prend une valeur proche de celle du mode de la distribution, l'aplatissement est faible.

L'aplatissement de la courbe des fréquences s'évalue par référence à la courbe des fréquences de la Loi Normale (ou loi de Laplace Gauss)

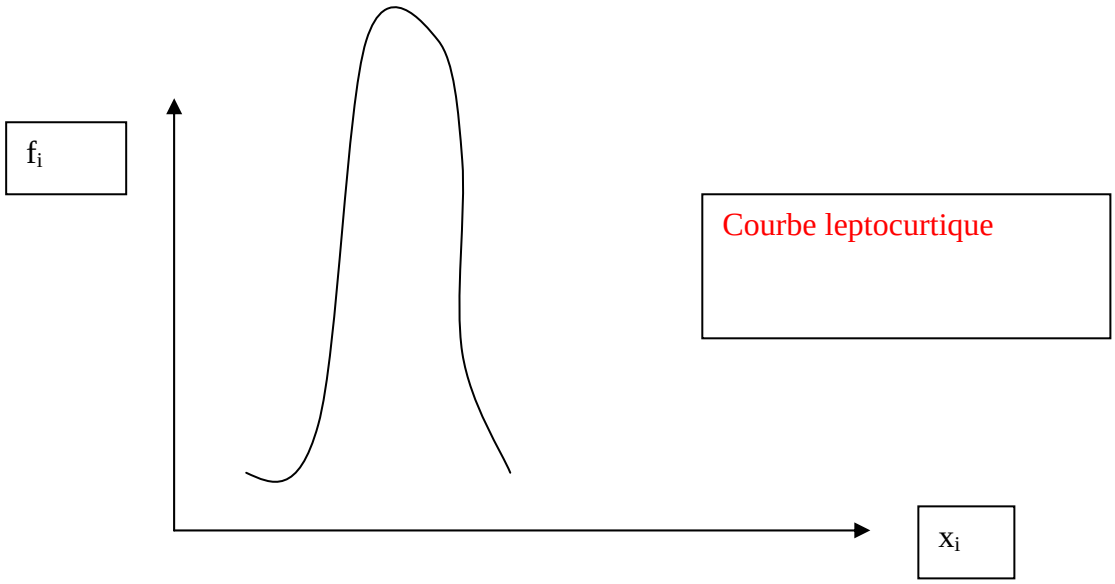


1.2.1 Définition

Une distribution est dite aplatie (ou platicurtique) si une forte variation de la variable entraîne une faible variation de la fréquence relative



Une distribution est leptocurtique, c'est-à-dire peu aplatie, si une faible variation de la variable *peut* entraîner une forte variation de la fréquence relative ou encore si une forte proportion des observations prend une valeur proche de celle du mode de la distribution



1.2.2 Les coefficients de l'aplatissement

1.2.2.1 Coefficient de Pearson

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma_x^4}$$

Pour la Loi normale : $\beta_2 = 3$
Plus la courbe est platicurtique et plus le coefficient β_2 est proche de 1
Plus la courbe est leptocurtique et plus le coefficient β_2 est grand (supérieur à 3)

1.2.2.2 Coefficient de Fisher

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma_x^4} - 3$$

Si $\gamma_2 = 0$ la distribution est normale
$\gamma_2 > 0$ la distribution est leptocurtique
$\gamma_2 < 0$ la distribution est platicurtique

2. La concentration

La notion de concentration est très importante en économie notamment en ce qui concerne les distributions de revenus ou de taille. Cette notion a trait à l'intensité du groupement des données et par conséquent elle est apparentée à celle de dispersion.
La concentration ne s'applique qu'aux distributions groupées prenant des valeurs positives, ou bien aux distributions dont les caractères sont susceptibles d'addition.
Deux méthodes peuvent être utilisées pour déterminer la concentration :

- détermination par le calcul
- détermination par les graphes

2.1 Détermination de la concentration par le calcul

Cette méthode consiste à mesurer la concentration en comparant l'écart entre la médiale (Ml) et la médiane (M_e) à l'intervalle de variation
4 étapes sont nécessaires à cette démarche :

2.1.1 Détermination de la médiane (M_e)

Etant donnée que les caractéristiques de concentration ne concernent que les distributions groupées prenant des valeurs positives, seul le cas des distributions groupées sera rappelé : la détermination de la médiane nécessite :

1) de repérer la classe médiane c'est-à-dire la classe contenant la valeur 50 pour la fréquence cumulée.
2) puis de calculer M_e par interpolation linéaire

2.1.2 Détermination de la médiale (M_l)

La médiale est une médiane que l'on calcule non plus sur les effectifs (n_i) de la série mais sur le produit $n_i.x_i$. La distribution étant groupée x_i représente en fait le centre de classe.
Le produit $n_i x_i$, représente par définition l'effectif multiplié par la valeur que prend la variable étudiée.

La médiale est donc la valeur du caractère x_i (centre de classe) qui partage la série $\{n_i.x_i ; x_i\}$ en 2 sous ensembles égaux. Par exemple, la médiale d'une distribution de salaires est la valeur du salaire qui partage la masse salariale en 2 sous ensembles égaux. Le salaire médial est tel que les salariés qui se situent en deçà gagnent autant que les salariés qui se situent au-delà.

La médiale se détermine à partir de la colonne des fréquences cumulées F(n.x) . Elle est toujours supérieure à la médiane puisque l'on raisonne en masse.

2.1.3 Ecart médiale-médiane

La médiale étant toujours supérieure à la médiane cet écart est forcément positif.

$$\Delta M = M_1 - M_e$$

2.1.4 Comparaison de ΔM à l'intervalle de variation

L'intervalle de variation est la différence entre la plus grande et la plus petite valeur du caractère.

Si ΔM est grand par rapport à l'intervalle de variation : la concentration est forte
Dans le cas des salaires, cela signifierait que l'inégalité entre les salaires est forte

Si ΔM est petit par rapport à l'intervalle de variation : la concentration est faible
Dans le cas des salaires, cela signifierait qu'il n'y a pas de grandes disparités salariales.

Si $\Delta M = 0$ c'est une situation d'équirépartition, ou d'absence de concentration.
Dans le cas des salaires, cela signifierait que tous les salariés touchent le même salaire.

2.2 La courbe de concentration

Cette analyse permet de construire la courbe de concentration (ou courbe de Lorenz) et de déterminer un indice de concentration (ou indice de Gini).

2.2.1 Construction de la courbe

Partons de l'exemple de la répartition des employés d'une PME selon le salaire mensuel en € .
Les observations étant groupées par classes, il nous faut adopter la convention du centre de classe

Modalités (€)	n_i	f_i **(%)**	**Centres de classe** x_i	F(x) **(%)**
[1000, 1250[	64	76,2	1125	76,2
[1250, 1500[	7	8,3	1375	84,5
[1500, 1750[	10	11,9	1625	96,4
[1750, 2000[	3	3,6	1875	100
Total	84	100		

2.2.1.1 Construction du tableau

Pour déterminer la concentration, nous avons besoin de connaître les masses salariales ($x_i.n_i$) et pas seulement sur le nombre de salaires (n_i) .

Complétons le tableau de la façon suivante :

Modalités (€)	n_i	f_i **(%)**	x_i	F(x) **(%)**	$n_i.x_i$ **(en €)**	$\frac{n_i.x_i}{\sum n_i.x_i}$	$F(n.x)$
[1000, 1250[	64	76,2	1125	76,2	72 000	69,55	69,55
[1250, 1500[	7	8,3	1375	84,5	9 625	9,3	78,85
[1500, 1750[	10	11,9	1625	96,4	16 250	15,7	94,55
[1750, 2000[	3	3,6	1875	100	5 625	5,45	100
Total	84	100			103 500	100	

La lecture de ce tableau montre que les salaires inférieurs à 1250 € représentent
76,2 % des salaires versés
et 69,55 % de la masse salariale

2.2.1.2 Construction du graphe

La courbe de concentration se construit dans un repère orthonormé en portant en abscisses les fréquences cumulées de la distribution statistique $\{n_i, x_i\}$, et en ordonnées les fréquences cumulées de la distribution $\{n_i.x_i, x_i\}$. La courbe que l'on obtient est appelée courbe de Lorenz
F(x) et F(n.x) sont des fréquences cumulées ce qui signifie qu'elles varient entre 0 et 1 (ou 0 et 100%).
Puisque F(x) et F(nx) s'annulent et sont égales à 1 en même temps, la courbe de concentration
-s'inscrit dans un carré dont la longueur des côtés est égale à l'unité
-passe par les 2 sommets 0B du carré. (concentration1.ppt)
-sa première bissectrice est telle que F(x)=F(n.x)

La courbe de concentration est toujours située au dessous de la diagonale OB du carré car $\forall i, F(x_i) > F(n_i x_i)$
Dans cette construction, la médiane apparaît en ordonnées et la médiale en abscisses.

Une situation d'équirépartition des données (F(x)=F(n.x)) correspond à l'absence de concentration (concentration2).
D'une façon générale :
Plus la concentration est faible, plus la courbe se rapproche de la diagonale 0B (concentration3)
Plus la concentration est forte, plus la courbe se rapproche des côtés du carré (concentration4). Autrement dit, plus la courbe de concentration s'éloigne de la bissectrice, plus la concentration est forte

D'où l'idée de caractériser la concentration par la surface comprise entre la bissectrice et la courbe de concentration : c'est précisément l'objet de l'indice de concentration

2.2.2 L'indice de GINI

Si on s'en tient à la courbe de Lorenz, la comparaison de la concentration dans deux populations différentes n'est pas facile, puisque ce qui est observé est un écart par rapport à la première bissectrice. Gini propose un ratio représentatif de l'écart entre la courbe de Lorenz et la première bissectrice.

L'aire maximale envisageable, i.e. l'écart maximal envisageable entre la courbe de Lorenz et la première bissectrice, est donnée par le triangle 0BA. L'indice permettant les comparaisons est celui qui rapporte l'aire comprise entre la courbe de Lorenz et la première bissectrice, appelée la ***surface de concentration***, à l'aire du triangle OBA (concentration5). C'est l'indice de Gini :

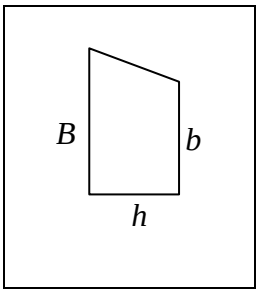
$$I_G = \frac{\text{aire de concentration}}{\text{aire du triangle OBA}}$$

La détermination de l'aire du triangle OBA ne pose pas de problème puisqu'il partage un carré dont les côtés mesurent 1 en deux. Son aire est : $\frac{1\times 1}{2} = 0{,}5$.

L'indice de Gini sera donc égal au double de l'aire comprise entre la courbe de concentration et la première bissectrice 0B. C'est un nombre sans dimension.

La détermination de la surface de concentration pose problème car nous ne disposons pas d'une formule pour la courbe de Lorenz permettant de calculer cette aire en passant par une intégrale. Cependant, sachant que cette courbe est une courbe brisée, nous pouvons estimer l'aire en calculant l'aire totale des trapèzes qui apparaissent sous la courbe. En faisant l'aire du triangle moins la somme des aires des trapèzes sous la courbe, nous obtenons la valeur de l'aire de la surface de concentration.

L'aire d'un trapèze est $S = \frac{(B+b)h}{2}$.



Avec $h = F(x_i) - F(x_{i-1})$, $B = F(n_i x_i)$ et $b = F(n_{i-1} x_{i-1})$

L'aire de la surface de concentration s'écrit :

$$\frac{1}{2} - \sum_i \frac{\left[F\left(n_{i-1}x_{i-1}\right) + F\left(n_i x_i\right)\right] \cdot \left[F\left(x_i\right) - F\left(x_{i-1}\right)\right]}{2} = \frac{1}{2} - \sum_i \frac{\left[F\left(n_{i-1}x_{i-1}\right) + F\left(n_i x_i\right)\right] \cdot f_i}{2}$$

Sachant que $I_G = \dfrac{\text{aire de concentration}}{\text{aire du triangle OBA}}$, et que l'aire du triangle OBA = $\dfrac{1}{2}$, on obtient :

$$I_G = 2 \cdot \left[\frac{1}{2} - \sum_i \frac{\left[F\left(n_{i-1}x_{i-1}\right) + F\left(n_i x_i\right)\right] \cdot f_i}{2}\right]$$

$$\text{ou } I_G = 1 - \sum_i \left[F\left(n_{i-1}x_{i-1}\right) + F\left(n_i x_i\right)\right] \cdot f_i$$

L'indice de Gini est donc égal au double de l'aire comprise entre la courbe de concentration et la première bissectrice 0B. C'est un nombre sans dimension qui varie entre 0 et 1 :

Si I_G tend vers 0 : La courbe de concentration est proche de la bissectrice 0B
La concentration est faible

Si I_G tend vers 1 : La courbe de concentration est proche des côtés du carré
La concentration est forte