RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieure et de La Recherche Scientifique

Université Saad Dahlab Blida

Faculté Des Sciences

Département de mathématiques

Option : Modélisation Stochastique et Statistique

Mémoire de Fin d'étude

Pour l'obtention du Diplôme de Master en Mathématique(LMD)

THEME

Inférence statistique de quelques indices en économie (indice de Gini et indice de Zenga)

Présenté par :

* Tifoura Salima
* Mellah Nabila

Promoteur : Mr. RASSOUL Abdelaziz

Soutenu devant le jury :

Mr TAMI Omar : Président

Mr LAIDI Mohamed M.A.A. ENST Alger : Examinateur

Promotion : octobre 2018

Remerciements

Louange à Dieu de m'avoir donnée le courage, la santé, la patience et la force de terminer mes études.

-Au terme de ce modeste travail, on tient à exprimer toute notre gratitude et tous nos remerciements à ceux qui nous ont aidés à réaliser ce mémoire, ainsi qu'à tous les enseignants du département de Mathématiques qui ont contribué à notre formation notamment pour leurs conseils et leurs dévouements.

-On tient à remercier très vivement Mr. RASSOUL .A d'avoir bien voulu nous encadré et suivre notre travail, on lui exprime notre reconnaissance pour son attention et précieux conseils contribuant à la réalisation de ce travail.

-Enfin, on remercie tous nos collègues et amis de l'université de Saad Dahlab Blida pour leurs soutiens moral et pour les moments merveilleux tout au long de notre cursus.

SALIMA & NABILA

Dédicace

Je remercie ALLAHle tout puissant de m'avoir donné le courage et la volonté de mener à ce terme ce présent travail.

Je dédie ce modeste travail et ma profonde gratitude à mon père décédé et ma mère pour l'éducation qu'ils m'ont prodigué, avec tous les moyens et au prix de toutes les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils mon enseigné depuis mon enfance.

A mes adorable frère DJILLALI et HICHEM et a chère sœur RAOUAE, et toute ma grande famille, pour leurs soutient et encouragements.

En particulier à mon binôme Salima et toute mes amies.

A toutes les personnes qui connaissent NABILA de prés ou de loi.

Nabila...

Dédicace

Je remercie ALLAH le tout puissant de m'avoir donné le courage et la volonté de mener à ce terme ce présent travail.

Je dédie ce modeste travail et ma profonde gratitude à mon père décédé et ma mère pour l'éducation qu'ils m'ont prodigué, avec tous les moyens et au prix de toutes les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils mon enseigné depuis mon enfance.

A' mes adorable frère MOUHAMED, AHMED et HICHEM et mes chères sœurs FAHIMA, SADJIA et KARIMA et les enfant RAMZI, LAMISS, Akram et Nour pour leurs patience, et toute ma grande famille, pour leurs soutient et encouragement.

En particulier à mon binôme Nabila et toute mes amies.

A toutes les personnes qui connaissent SALIMA de prés ou de loin.

Salima...

Table des matières

Table des figures

Liste des tableaux

ملخص

تتناول هذه الرسالة مشكلة الاستدلال الاقتصادي ومناقشات مؤشرات التفاوت في الدخل الأكثر استخدامًا ، ومؤشر Gini ومؤشر Zenga. يتم عرض خصائصه وجمعياته بواسطة أداة أخرى تستخدم على نطاق واسع لتمثيل اللامساواة: فهو منحنى لورينز. كما يناقش بطريقة موسعة هذه المؤشرات تستند إلى نماذج خطرة مثل باريتو ، فريشيت ، .... يوضح معامل المقدرات التجريبية أنها غير كافية للتوزيعات ذات الذيل الثقيل ، لأن ذلك يقترح شبه مقدرات. استنادا إلى المقاييس القصوى ، فقد أظهرنا الخصائص المقاربة لهذين المقدرين ، ثم يتم عرض بعض الرسوم التوضيحية من خلال بعض نتائج المحاكاة.

0.1 Résumé

Ce mémoire traite le problème de l'inference économique et des discussions sur les indicateurs d'inégalité des revenus les plus couramment utilisés, l'indice de Gini et l'indice de Zenga. Ses propriétés et associations sont présentées par un autre outil largement utilisé pour représenter l'inégalité : c'est la courbe de Lorenz. Il discute également d'une façon élargie de ces indicateurs basant sur des modèles dangereux telle que Pareto, Fréchet,.... Le traitement des estimateurs empiriques montre qu'ils sont insuffisants pour les distributions à queues lourdes. Pour cela on propose des estimateurs semi paramétriques basés sur des quantiles extrêmes, on a montré les propriétés asymptôtiques de ces estimateurs, ensuite des illustrations sont présentées par quelques résultats de simulations.

0.2 Abstract

This thesis addresses the problem of economic inference and discussions of the most commonly used income inequality indicators, the Gini index and the Zenga index. Its properties and associations are presented by another tool widely used to represent inequality : it is the Lorenz curve. It also discusses in an enlarged way these indicators based on dangerous models such as Pareto, Fréchet, ... The treatment of the empirical estimators shows that they are insufficient for the distributions with heavy tails, for that one proposes semi estimators parametric based on extreme quantiles, we have shown the asymptotic properties of these estimators, then some illustrations are presented by some simulation results.

Introduction générale

0.3 Introduction générale

En économie, un indicateur est une statistique construite afin de mesurer certaines dimensions de l'activité économique, ceci de façon aussi objective que possible. Leurs évolutions ainsi que leurs corrélations avec d'autres grandeurs sont fréquemment analysées à l'aide de méthodes économétriques.

Les indicateurs sont construits par l'agrégation d'indices qui figurent dans un document appelé « tableau de bord ». La construction des indicateurs découle d'un choix de conventions qui traduisent plus ou moins bien certaines priorités et valeurs éthiques et morales. Le « Tableau économique » de François Quesnay, l'un des premiers physiocrates qui a vécu au XVIIIe siècle, constitue l'un des premiers exemples d'un tel indicateur visant à mesurer la richesse d'un pays. Depuis les développements des comptes nationaux après la Seconde Guerre mondiale, le produit intérieur brut (PIB) et le produit national brut (PNB) sont les indicateurs les plus courants. Par ailleurs, il existe d'autres indicateurs qui prennent en compte d'autres facteurs ignorés par le PNB et le PIB afin de mesurer le bien-être des habitants d'un pays ; en incluant par exemple des indicateurs de santé, d'espérance de vie, de taux d'alphabétisation. Le Programme des Nations Unies pour le développement (PNUD) a ainsi créé l'indice de développement humain (IDH) dans les années 1990.

Des tentatives pour prendre en compte d'autres dimensions telles la sécurité ou pour inclure la « soutenabilité écologique » de l'activité économique dans des indicateurs ont aussi été menées plus récemment.

L'inégalité est un concept plus large que la pauvreté dans la mesure où elle est définie sur l'ensemble de la population et ne se concentre pas uniquement sur les pauvres. La mesure la plus simple de l'inégalité trie la population du plus pauvre au plus riche et montre le pourcentage de la dépense (ou du revenu) attribuable à chaque cinquième (quintile) ou dixième (décile) de la population. Le quintile le plus pauvre représente généralement de 6 à 10% de toutes les dépenses, le quintile supérieur étant de 35 à 50%. Une mesure populaire de l'inégalité est L'indice de Gini, basé sur le travail de Gini, est utilisé pour décrire l'inégalité de revenu dans une population.
Les pays publient leur propre indice de Gini. De nombreuses institutions, compris la Banque mondiale et la CIA, calculent l'indice de Gini des pays du monde. L'inégalité des revenus dans les «limites optimales» favorise la croissance. Le taux d'inégalité des revenus dans les pays riches du monde évite l'égalitarisme extrême et l'extrême inégalité. Il n'y a pas de corrélation entre Gini et la richesse pour les pays les plus pauvres, qui va de 0 (égalité parfaite) à 1 (inégalité parfaite), mais se situe généralement entre 0,3 et 0,5 pour les dépenses par habitant. L'indice de Gini est dérivé de la courbe de

loi Pareto et Fréchet respectivement.

Chapitre 1: Rappel sur les variables aléatoires

Chapitre 1

Rappel sur les variables aléatoires

1.1 Introduction

Ce chapitre introduit les concepts essentiell des modèles probabilistes afin d'aborder l'inférence statistique : definition d'une variable aléatoire, les variable discrêtes ou continues, aprés avoir les lois plus utilisées sont décrites : Uniforme discrête, Bernouli, Binomiale et Poisson, et des variables continues : Exponentielle, Normal, Pareto, Uniforme, Burr et de Fréchet. ensuit l'échantillonnge et les méthodes d'estimations paramétriques et non paramétriques.

1.2 Définition d'une variable aléatoire

Considérons un ensemble fondamental E correspondant à une certaine expérince. Les éléments de E, résultats possibles de l'expérience , ne sont généralement pas des nombres. Il est cependant utile de faire correspondre un nombre à chaque élément de E, en vue de faire ensuite des calculs.

Pour un jet de dé, il semble naturel de faire correspondre à la face obtenue par le jet, le nombre de points qu'elle porte, mais ce n'est pas une obligation. Si on jette 2 de dés, on s'intéressera par exemple à la somme des poits obtenus. Pour une carte à jouer, il faut convenir d'une valeur carte à jouer.
Un variable aléatoire X sur un ensemble fondamental E, est une application de E dans $\mathbb{R}$: à tout résultat possible de l'expérience (à tout élément de E), la variable aléatoire X fait correspondre un nombre. Lorsque E est fini ou infini dénomble, toute application de E dans $\mathbb{R}$ est une variable aléatoire. Lorsque E est non dénombrable. il existe certaines applications de E dans $\mathbb{R}$ qui ne sont pas des variables aléatoires. En effet, la définition rigoureuse

d'une variable aléatoire X impose que tout intervalle de $\mathbb{R}$ soitl'image d'un événement de E par l'application X. Cette condition est vérifiée pour tout application X si E est fini ou dénombrable, puisque toute partie de E est un événement. Ce n'est plus vrai si E est non dénombrable. Heureusement, les applications choisies naturellement sont des variables aléatoires. On parle de variable aléatoire discrète lorsque la variable est une application de E dans un sous ensemble discret de $\mathbb{R}$, le plus souvent $\mathbb{N}$ ou une partie de $\mathbb{N}$. On parle si non de variable aléatoire continue.

Pour un nombre réel a donné, l'événement constitué de tous les résultats ω d'expérience teles que $X(\omega) = a$ est noté $[X(\omega) = a]$, ou, en abrégé, $X = a$.

Pour deux nombres réels a et b $(a \leq b)$, l'événement constitué de tous les résultats ω d'epérience tels que $a \leq X(\omega) \leq b$ est noté $[a \leq X(\omega) \leq b]$ ou , en abrégé, $a \leq X \leq b$.

Si X est Y sont des variables aléatoires définies sur le même ensemble fondamental E , et si k est une constante, on peut montrer que les fonctions suivantes sont aussi des variables aléatoires :

- $(X + Y)(\omega) = X(\omega) + Y(\omega)$
- $(X + k)(\omega) = X(\omega) + k$
- $(kX)(\omega) = kX(\omega)$
- $(XY)(\omega) = X(\omega)Y(\omega)$

pour tout élément ω de E.

1.2.1 Variable aléatoire discrète

Définition 1.1 *Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé. On appelle variable aléatoire discrète sur $(\Omega, \mathcal{F}, P)$ toute application X,*

$$\begin{array}{rcl} X & : & \Omega \to \mathbb{R} \\ \omega & \mapsto & X(\omega) \end{array}$$

vérifiant les deux conditions :

1. L'ensemble des images $X(\Omega) = \{X(\omega), \omega \in \Omega\}$ est une partie au plus dénombrable de $\mathbb{R}$. On peut donc numéroter ses éléments par des indices entiers
$$X(\Omega) = \{x_0, x_1, ..., x_k, ...\}.$$
2. Pour tout $x_k \in X(\Omega), A_k = \{\omega \in \Omega, X(\omega) = x_k\}$ fait partie de la famille $\mathcal{F}$ d'événements aux quels on peut attribuer une probabilité par P. L'événement A_k est aussi noté $X^{-1}(\{x_k\})$ ou plus commodément $\{X = x_k\}$. Nous utiliserons l'abréviation $v.a.$ pour variable aléatoire.

Remarquons que la famille de tous les A_k forme une partition de Ω : on classe chaque élément de Ω selon son image par X. Il en résult :

$$\sum_{x_k \in X(\Omega)} P(A_k) = \sum_{x_k \in X(\Omega)} P(X = x_k) = 1 \tag{1.1}$$

Dans cette écriture, les sommes sont des séries convergentes si $X(\Omega)$ est infini et des sommes ordinaires lorsque l'ensemble $X(\Omega)$ est fini.

Loi d'une variable aléatoire discrète

L'application X permet de transporter la probabilité P de Ω sur $\mathbb{R}$:on considère les $P(X = x_k)$ comme des masses ponctuelles situe elles aux points x_k de la droite réelle. La probabilité d'une partie quelconque de $\mathbb{R}$ est alors définie comme la somme de ses masses ponctuelles.

Définition 1.2 *Soit X une variable aléatoire discrète sur $(\Omega, \mathcal{F}, P)$. On lui associe la fonction d'ensemble P_X définie sur la famille de toutes les parties de $\mathbb{R}$ en posant :*

$$p_k = P_X(\{x_k\}) = P(A_k) = P(X = x_k) \tag{1.2}$$

puis pour tout $B \subset \mathbb{R}$:

$$P_X(B) = \sum_{x_k \in B} P(X = x_k) = \sum_{x_k \in B} p_k \tag{1.3}$$

La fonction d'ensembles P_X ainsi définie est une probabilité sur la famille $\mathcal{P}(\mathbb{R})$. de toutes les parties de $\mathbb{R}$.

Remarquons que la définition de $P_X(B)$ a toujours un sens en convenant qu'une somme indexée par l'ensemble vide vaut 0 et en notant que s'il y a une infinité de x_k dans B, la somme sur B des p_k est une sous-série de la série à termes positifs convergente : $\sum_{x_k \in X(\Omega)} p_k = 1$.

Remarque 1.1 *Deux variables aléatoires peuvent avoir a même loi sans être égales. Par exemple considérons le jet de deux dés, l'un bleu et l'autre rouge.*

Notons X le nombre de points indiqué par le dé bleu et Y celui du rouge. Les variables aléatoires X et Y sont définies sur le même espace probabilisé $\Omega = \{1,2,3,4,5,6\}^2$ muni de l'équiprobabilité. On a $X(\Omega) = Y(\Omega) = \{1,2,3,4,5,6\}$ et :

$$\forall k \in \{1,2,3,4,5,6\}, P(X=k) = \frac{1}{6}, P(Y=k) = \frac{1}{6}.$$

Donc X et$Y$ ont même loi : $P_X = P_Y$. pour autant on n'a pas l'égalité des variables aléatoires X et Y qui signifierait $X(\omega) = Y(\omega)$pour tout $\omega \in \Omega$ (égalité de deux applications). Autrement dit, en lançant deux dés on obtiendrait à coup sûr un double. Par contre nous pouvons considérer l'événment $\{X = Y\}$ dont la réalisation n'est pas certaine et calculer sa probabilité :

$$P(X=Y) = P\left(\bigcup_{k=1}^{6} \{(X,Y) = (k,k)\}\right) = \frac{6}{36} = \frac{1}{6} \tag{1.4}$$

On en déduit :$P(X \neq Y) = 5/6$.

Fonction de répartition

Définition 1.3 *On appelle fonction de répartition de la variable aléatoire X, la fonction F_X définie sur $\mathbb{R}$ par :*

$$\forall x \in \mathbb{R}, F_X(x) = P_X(]-\infty, x[) = P(X \leq x) \tag{1.5}$$

on a aussi :

$$F_X(x) = \sum_{\substack{x_k \in X(\Omega) \\ X_k \leq X}} P(X = x_k). \tag{1.6}$$

Théorème 1.1 *La fonction de répartition F_X d'une variable aléatoire discrète X est croissante sur $\mathbb{R}$, continue à droite et limitée à gauche en tout point, elle tend vers 0 en $-\infty$ et vers 1 en $+\infty$. Elle caractérise la loi de X, autrement dit $F_X = F_Y$ si et seulement si les variables aléatoires X et Y ont même loi.*

Lois discrètes classiques

Loi Uniforme

Définition 1.4 *On dit qu'une variable aléatoire X suit une* ***loi uniforme discrète*** *lorsqu'elle prend ses valeurs dans $\{1,...,n\}$ avec des probabilités élémentaires identiques. Puisque la somme des ces dernières doit valeur 1, on en déduit qu'elles doivent toutes être égales à un $1/n$:*

$$\forall k = 1...n, \mathbb{P}[X=k] = \frac{1}{n}. \tag{1.7}$$

Proposition 1.1 *(Espérance et variance) : On calcule aisément*

$$\mathbb{E}[X] = \frac{n+1}{2}$$

$$V(X) = \frac{n^2-1}{12}$$

Preuve:

$$\begin{aligned} \mathbb{E}[X] &= 1.\frac{1}{n} + 2.\frac{1}{n} + 3.\frac{1}{n} + ... + n.\frac{1}{n} \\ &= \frac{1}{n}.\sum_{k=1}^{n} k = \frac{1}{n}.\frac{n(n+1)}{2} = \frac{n+1}{2} \end{aligned} \tag{1.8}$$

$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ est la somme des premiers termes d'une suite arithmétique de raison 1 de premier terme 1.

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= 1^2.\frac{1}{n} + 2^2.\frac{1}{n} + 3^2.\frac{1}{n} + ... + n^2.\frac{1}{n} \\ &= \frac{1}{n}.\sum_{k=1}^{n} k^2 = \frac{1}{n}.\frac{n(n+1)(2n+1)}{6} \\ &= \frac{(n+1)(2n+1)}{6}. \end{aligned} \tag{1.9}$$

$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$ est un résultat classique qui se démontre par récurrence.
Ainsi,

$$\begin{aligned} V[X] &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= (n+1)\left[\frac{2n+1}{6} - \frac{n+1}{4}\right] \\ &= (n+1)\left[\frac{4n+2-3n-3}{12}\right] \\ &= (n+1)\frac{n-1}{12} = \frac{n^2-1}{12} \end{aligned}$$

■

Loi de Bernoulli

Définition 1.5 *Cette loi est celle de toute variable aléatoire X modélisant une expérience dont l'issue ne possède que deux alternatives de type "succès ou échec", "vrai ou faux", "marche ou arrêt", "pile ou face", etc. Un succès est représenté par l'évènement $\{X = 1\}$ tandis que $\{X = 0\}$ correspond à un échec $X(\Omega) = \{0; 1\}$. Puisque l'on a $\mathbb{P}[X = 0] = 1 - \mathbb{P}[X = 1]$, la loi de X ne dépend que d'un paramètre (la probabilité de succès) ; on parle alors de la loi de Bernoulli de paramètre p caractérisée par*

$$X = \begin{cases} p & si\ X = 1 \\ 1 - p & sinon\ X = 0 \end{cases} \tag{1.10}$$

Exemple 1.1 *On joue au Pile au Face avec proba p de tomber sur Pile (et donc $1 - p$ de tomber sur Face).*

Soit

$$X = \begin{cases} 1 & \text{si on obtient Pile} \\ 0 & \text{si non} \end{cases} \tag{1.11}$$

alors X suit une Bernoulli(p).

Paramétres d'une Bernoulli

Proposition 1.2 *(Espérence et variance)*

$$\mathbb{E}[X] = p \tag{1.12}$$

$$V[X] = p(1 - p) \tag{1.13}$$

Loi Binomiale

Définition 1.6 *La **loi binomiale** est la loi de probabilité d'une variable aléatoire représentant une série d'épreuves de **Bernoulli** possédant les propriétés suivantes :*

1. *Chaque épreuve donne lieu à deux éventualités exclusives de probabilités constantes p et $q = 1 - p$.*
2. *Les épreuves répétées sont indépendantes les unes des autres.*
3. *La variable aléatoire X correspondante prend pour valeur le nombre de succès dans une suite de n épreuves.*

Cette loi est donc caractérisée par deux paramètre : n et p.

Lorsque d'une telle expérience, on dit que X suit une binomiale $\mathcal{B}(n,p)$, à valeurs dans $X(\Omega) = \{0, 1, 2, ..., n\}$.

Exemple 1.2 *On joue n fois au pile ou face.*
Pour $1 \leq i \leq n$, *on pose*

$$X_i = \begin{cases} 1 & si\ on\ obtient\ Pile \\ 0 & si\ non \end{cases} \quad au^{i\acute{e}me} lancés.$$

Soit Y *le nombre de "Piles" obtenus au cours des* n *lancers indépendants de la pièce. Alors,*

$$Y = X_1 + X_2 + ... + X_n$$

et Y *suit une loi binomiale* $\mathcal{B}(n,p)$.

Proposition 1.3 *Les probabilités élémentaires d'une variable aléatoire* X *suivant une loi binomiale* $\mathcal{B}(n,p)$ *sont données pour tout nombre de succès* $k = 1...n$ *par :*

$$\mathbb{P}[X = k] = C_n^k.p^k.(1-p)^{n-k}. \tag{1.14}$$

Remarque 1.2 *On a bien, en utilisant la formule du binôme,*

$$\sum_{k=0}^{n} \mathbb{P}[X = k] = \sum_{k=0}^{n} C_n^k.p^k.(1-p)^{n-k} = 1.$$

Proposition 1.4 *(Espérance et variance)*

$$\mathbb{E}[X] = np \tag{1.15}$$

$$V[X] = np(1-p). \tag{1.16}$$

Preuve: On a l'écriture $X = X_1 + X_2 + ... + X_k + ... + X_n$ ou les X_k sont n variables aléatoires de Bernoulli indépendantes. En effet, par linéarité de l'espérance

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + ... + \mathbb{E}[X_k] + ... + \mathbb{E}[X_n] = n.\mathbb{E}[X_1] = n.p.$$

et par indépendance des variables aléatoires $(X_k)_{k=1...n}$

$$V[X] = V[X_1] + V[X_2] + ... + V[X_k] + ... + V[X_n] = n.V[X_1] = n.p(1-p).$$

■

Loi de Poisson Cette loi peut modéliser les événement s rares. Par exemple elle peut modèliser le nombre d'appels reçus par un standard téléphonique, le nombre de voyageurs se présentant à un guichet dans la journée,... etc. Pour ces raisons, elle s'exprime à l'aide de la fonction exponentielle et dépend d'un paramètre $\lambda > 0$, qui correspond au nombre moyen d'occurence du phénomène observé pendant la durée.

Plus formellement :

Définition 1.7 *Une variable aléatoire X suit une* ***loi de Poisson de paramêtre*** $\lambda > 0$, *notée* $\mathcal{P}(\lambda)$ *lorsque* $X(\Omega) = \mathbb{N}$ *et pour tout* $k \in \mathbb{N}$

$$\mathbb{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0, 1, 2, ... \tag{1.17}$$

Notation 1.1 $X \rightsquigarrow \mathcal{P}(\lambda)$

Paramètre d'une loi de poisson

Proposition 1.5 *(Espérance et variance)*

$$\mathbb{E}[X] = V[X] = \lambda. \tag{1.18}$$

Exemple 1.3 *Si on sait qu'en général un standard téléphonique reçoit 20 appeles dans la journée et que l'on peut modéliser le nombre aléatoire d'appels par une loi de Poissson, on pourra calculer la probabilité d'avoir k appels, pour tout k, à l'aide des formules données par une loi de Poisson $\mathcal{P}(20)$.*

Remarque 1.3 *Dans la pratique, des tables donnant les probabilités élémentaires pour différentes valeurs du paramètre sont dispoonibles et peuveut être utilisées.*

Approximotion d'une binomiale par une poisson

Proposition 1.6 *Soit X_n des v.a telleque $X_n \rightsquigarrow \mathcal{B}(n, p_n)$ et $Y \rightsquigarrow \mathcal{P}(\lambda)$ alors pour tout $0 \leq k \leq n$, on a : si $\lim_{n\to+\infty} np_n = \lambda$, alors*

$$\lim_{n\to+\infty} \mathbb{P}(X_n = k) = \mathbb{P}(Y = k)$$

On dit que X_n converge en loi vers la loi de Poisson et on écrit :

$$X_n \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} Y. \tag{1.19}$$

Preuve: On a :

$$\begin{aligned} \mathbb{P}(X_n = k) &= C_n^k p^k (1-p)^{n-k} \\ &= \frac{n(n-1)...(n-k+1)}{k!} \frac{p_n^k}{(1-p_n)^k} (1-p_n)^n \\ &\simeq \frac{n(n-1)...(n-k+1)}{n^k} \frac{\lambda^k}{k!} \frac{1}{\left(1-\frac{\lambda}{n}\right)^k} \left(1-\frac{\lambda}{n}\right)^n \\ &\simeq \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned} \tag{1.20}$$

■

1.2.2 Variable aleatoire continue

Une variable aléatoire X dont l'ensemble image $X(E)$ est un intervalle de $\mathbb{R}$ est une variable aléatoire continue. Rappelons que, par définition d'une variable aléatoire, $(a \leq X \leq b)$ est un événment de X dont la probabilité est bien définie. on définit la loi de probabilitié de X, ou distribution de X, à l'aide d'une fonction $f(x)$, applée densité de probabilité de X, telle que

$$\int_a^b f(x)\,dx = \Pr(a \leq X \leq b). \tag{1.21}$$

1. Si f est donnée, la probabilité $\Pr(a \leq X \leq b)$ est la surface sous la courbe et entre a et b

1. Le passage du discrèt au continu transforme les sommes $\sum$ en intégrales $\int$ et p_i en $f(x)\,dx$.

Ainsi, soit X une variable aléatoire discrète et p_i sa distribution.
La formule

$$\Pr(x_k \leq X \leq x_n) = \sum_{i=k}^{n} P_i \tag{1.22}$$

En utillisant cette analogie, on admettra les définitions suivantes pour une variable aléatoire X continue de distribution $f(x)$

1. $f(x) \geq 0$ (analogue à $p_i \geq 0$)
2. $\int_{\mathbb{R}} f(x)\,dx = 1$ (analogue à $\sum_i p_i = 1$)

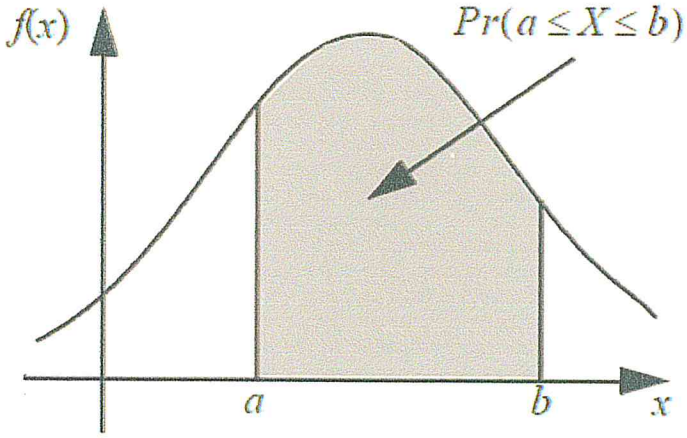


FIG. 1.1 – La probabilité d'une variable aléatoire dans un interval [a,b]

3. $\mu_x = E(x) = \int_{\mathbb{R}} xf(x)\,dx$ (analogue à $\sum_i x_i p_i$)
4. $\sigma_x^2 = var(x) = \int_{\mathbb{R}} (x-\mu_x)^2 f(x)\,dx$ (analogue à $\sum_i (x_i-\mu_x)^2 p_i$)
5. $\sigma_x^2 = var(x) = \int_{\mathbb{R}} x^2 f(x)\,dx - \mu_x^2$ (analogue à $\sum_i x_i^2 p_i - \mu_x^2$)
6. $\sigma(x) = \sigma_x = \sqrt{var(x)}$
7. $F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(\tau)\,d\tau$ (analogue à $\sum_{x_i \leq x} p_i$)
8. $\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$

Cet exemple montre la densitè de probabilité et la fonction de répartition d'une certaine variable aléatoire continue.

La probabilité sur l'intervalle $[a\ ,b]$ est la surface sous la courbe de densité limitée par cet intervalle, c'est aussi la différence des hauteurs $F(b) - F(a)$ si on utilise la fonction de répartition. Contrairement au cas des variables discrètes la fonction de répartition est ici continue.

Pour résumer la différence entre le cas discrèt et le cas continu, un point dans le cas discrèt correspond à un intervalle dans le cas continu, la somme discrète correspond à l'intégrale.

Loi exponentielle

Définition 1.8 *Une **loi exponentielle** modélise la durée de vie d'un phénomène sans mémoire, ou sans vieillissement, ou sans usure : la probabilité*

FIG. 1.2 – Diagramme en batôn d'une v.a. discète

que le phénomène dure au moins $s+t$ heures sachant qu'il a déjà duré t heures sera la même que la probabilité de durer s heures à partir de sa mise en fonction initiale. En d'autres termes, le fait que le phénomène ait duré pendant t heures ne change rien à son espérance de vie à partir du temps t. Plus formellement, soit X une variable aléatoire définissant la durée de vie d'un phénomène, d'espérance mathématique $\mathbb{E}(X)$. On suppose que :

$$\forall (s,t) \in \mathbb{R}_+^2, P_{X>t}(X > s+t) = P(X > s).$$

Alors, la densité de probabilité de X est définie par :

$$f(t) = \begin{cases} 0 & si\ t < 0 \\ \frac{1}{\mathbb{E}(X)} e^{-\frac{t}{\mathbb{E}(X)}} & pour\ tout\ \ t \geq 0 \end{cases} \tag{1.23}$$

et on dit que X suit une **loi exponentielle de paramètre** (ou de facteur d'échelle) $\lambda = \frac{1}{\mathbb{E}(X)}$. Réciproquement, une variable aléatoire ayant cette loi vérifie la propriété d'être sans mémoire.

Cette loi permet entre autres de modéliser la durée de vie de la radioactivité ou d'un composant électronique. Elle peut aussi être utilisée pour décrire par exemple le temps écoulé entre deux coups de téléphone reçus au bureau, ou le temps écoulé entre deux accidents de voiture dans lequel un individu donné est impliqué.

Densité de probabilité

La densité de probabilité de la distribution de paramètre $\lambda > 0$ prend la forme :

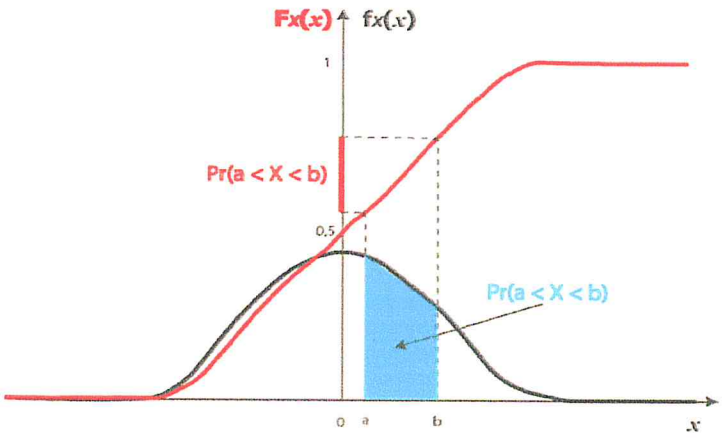


FIG. 1.3 – Densité de probabilité (en noire) et la fonction de répartition d'une v. a. continue

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{1.24}$$

La distribution a pour support l'intervalle $[0, +\infty[$.

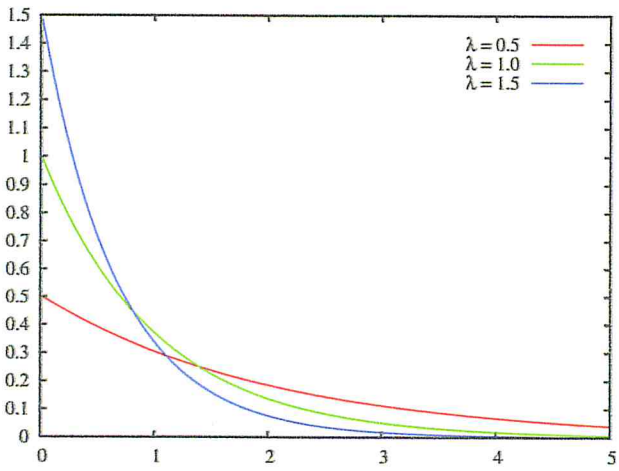


FIG. 1.4 – Densité de probabilité de la lois exponentielle

Fonction de répartition

La fonction de répartition est donnée par :

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{1.25}$$

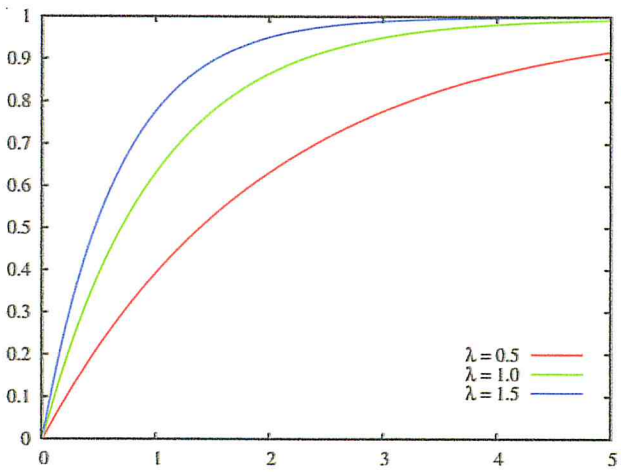


FIG. 1.5 – foncion de répartition de loi exponentielle

Espérance, variance, écart_type, médiane

Soit X une variable aléatoire qui suit une loi exponentielle de paramètre λ.

Nous savons, par construction, que l'espérance mathématique de X est $\mathbb{E}(X) = \frac{1}{\lambda}$.

On calcule la variance en intégrant par parties ; on obtient : $V(X) = \frac{1}{\lambda^2}$.

L'écart type est donc $\delta(X) = \frac{1}{\lambda}$.

La médiane, c'est-à-dire le temps T tel que $\mathbb{P}(X > T) = 0,5$, est $m = \frac{\ln(2)}{\lambda} = \mathbb{E}(X)\ln(2)$.

Le fait que la durée de vie soit sans vieillissement se traduit par l'égalité suivante :

$$\forall T \geq 0, \quad \mathbb{P}_{X>T}(X > T + t) = \mathbb{P}(X > t), \tag{1.26}$$

Par **Théorème de Bayes** on a :

$$\mathbb{P}_{X>T}(X > T + t) = \frac{\mathbb{P}(X > T \ \ et \ \ X > T + t)}{\mathbb{P}(X > T)} = \frac{\mathbb{P}(X > T + t)}{\mathbb{P}(X > T)}. \tag{1.27}$$

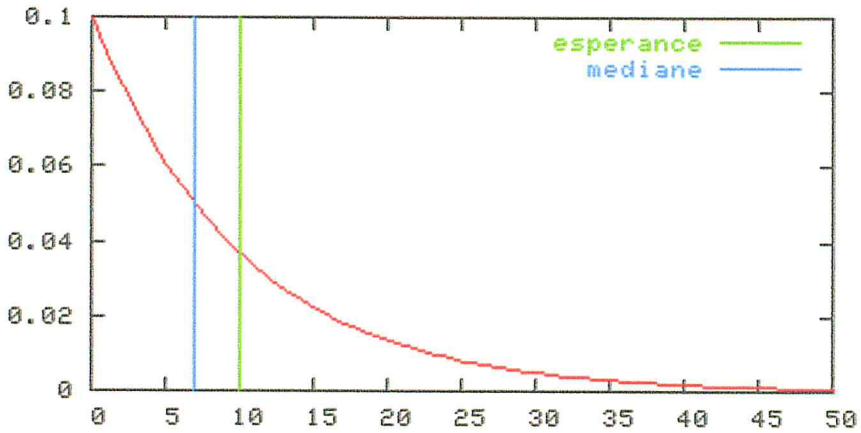


FIG. 1.6 – Espérance et variance d'une v.a. exponontielle

En posant $\mathbb{P}(X > t) = 1 - F(t) = G(t)$, la probabilité que la durée de vie soit supérieure à t, on trouve donc :

$$\frac{G(T+t)}{G(T)} = G(t). \tag{1.28}$$

Puisque la fonction G est monotone et bornée, cette équation implique que G est une fonction **exponentielle**. Il existe donc k réel tel que pour tout t :

$$G(t) = e^{kt}.$$

Notons que k est **négatif,** puisque G est inférieure à 1. La densité de probabilité f est définie, pour tout $t \geq 0$, par :

$$f(t) = -ke^{kt}$$

Le calcul de l'espérance de X, qui doit valoir $\mathbb{E}(X)$ conduit à l'équation :

$$\int_0^{+\infty} -kte^{kt}dt = \mathbb{E}(X)$$

On calcule l'intégrale en intégrant par parties ; on obtient :

$$k = -\frac{1}{\mathbb{E}(X)} = -\lambda$$

Donc

$$\mathbb{P}(X > t) = e^{-\frac{t}{\mathbb{E}(X)}}$$

et $f(t) = \frac{1}{\mathbb{E}(X)} e^{-\frac{t}{\mathbb{E}(X)}}$.

Loi normale

Définition 1.9 *La distribution normale. ou de Laplace – Gauss, appelée aussi gaussienne, est une distribution continue qui dépend de deux paramètres μ et σ.On la note $N(\mu, \sigma^2)$. Le paramètre μ peut être quelconque mais σ est positif. Cette distrribution est définie par :*

$$f(x : \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{1.29}$$

C'est une des lois les plus importantes comme vous le verrez à l'occasion du théorème central limite.

Propriétés :

Allure de courbe :

La loi normale, notée $\mathcal{N}(\mu, \sigma^2)$, est symétrique par rapport à la droite d'abscisse μ.

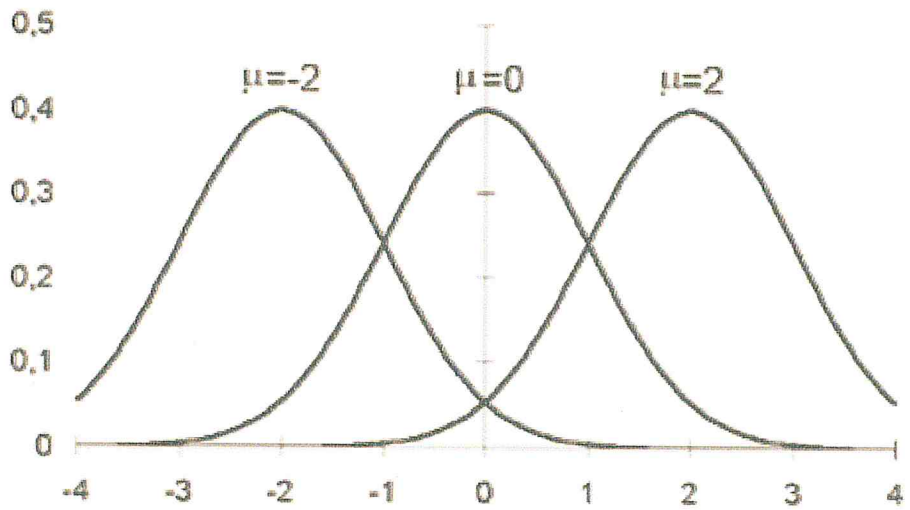
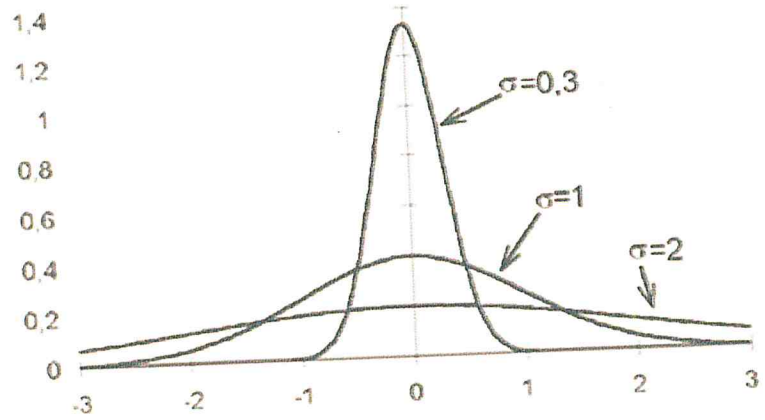


FIG. 1.7 – Lois normale $\mathcal{N}(\mu, 1)$ pour les valeurs de $\mu = -2; 0$ et 2.



$\mathcal{N}(0,\sigma^2) pour les valeurs de \sigma = 0.3, 1$ et 2.

Caractéristiques de loi normale

Espérance	μ
Variance	σ^2
Ecart-type	σ

La distribution normale centrée réduite :

On dit que la distribution est centrée si son espérance μ est nulle : elle est dite réduite si sa variance σ^2 (et son écart-type) est égale à 1. La distribution normale centrée réduite $\mathcal{N}(0,1)$ est donc définie par la formule

$$f(t:0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

Loi de Pareto

La loi de Pareto s'applique pour les distributions tronquées. Prenons un exemple de la vie courante, en France, la borne basse du salaire horaire est forcément le SMIG, il ne peut pas en être autrement. La loi de Pareto permet de tenir compte de cette contrainte en restreignant le domaine de définition de la v.a. X .

Définition 1.10 *La loi possède 2 paramètres, $\alpha > 0$ et c qui introduit la contrainte $x > c$. Le domaine de définition de X est $]c;+\infty[$. La fonction de densité est monotone décroissante, elle s'écrit*

$$f(x) = \frac{\alpha}{c}\left(\frac{c}{x}\right)^{\alpha+1}, x > c. \tag{1.30}$$

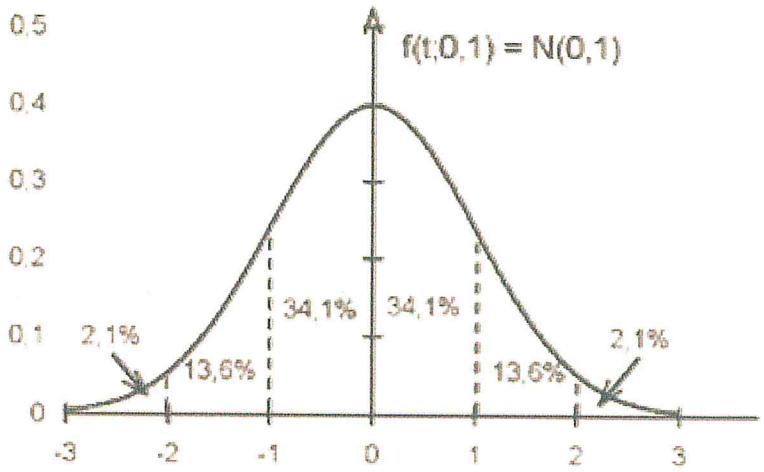


FIG. 1.8 – Loi normale centrée réduite $\mathcal{N}(0,1)$

La fonction de répartition est directement obtenue avec

$$F(x) = 1 - \left(\frac{c}{x}\right)^{\alpha} \tag{1.31}$$

Caractéristiques de la loi

$$E(X) = \frac{\alpha}{\alpha - 1}, \quad \text{pour} \quad \alpha > 1 \tag{1.32}$$

$$V(X) = \frac{\alpha}{(\alpha - 1)^2 (\alpha - 2)} c^2, \text{ pour } \alpha > 2 \tag{1.33}$$

$$E[X^k] = \frac{\alpha}{\alpha - c}, \text{ pour } \alpha > c \tag{1.34}$$

Loi uniforme

Description

La principale caractéristique de la loi unforme continue est que la probabilité d'être dans intravalle dépend uniquement de la largeur de l'intervalle et non de la position de lintervalle dans le domaine de définition. La fonction de densité prend une forme rectangulaire. La loi est symétrique.

Remarque 1.4 *(Loi discrète uniforme) On peut définir le pendant de la loi uniforme pour les v.a. discrètes. Il s'agit dans ce cas d'un cas particulier de la loi multinomaile où les états E_k sont équiprobables. La loi repose sur deux*

parametres a et b qui sont respectivement la borne basse et la borne haute du domaine de définition de la v.a.X. La fonction de densité est alors :

$$f(x) = \begin{cases} \frac{1}{a-b} & si\ a \leq x \leq b \\ 0 & sinon \end{cases} \tag{1.35}$$

Dans le cas de la loi uniforme, il est aisé de caluler la fonction de répartition :

$$f(x) = \frac{x-a}{b-a} \tag{1.36}$$

Caractéristiques de la loi

$$E(X) = \frac{a+b}{2} \tag{1.37}$$

$$V(X) = \frac{(b-a)^2}{12} \tag{1.38}$$

Loi de Burr

En théorie des probabilités, en statistique et en économétrie, la **loi de Burr**, loi de Burr de type XII, loi de Singh-Maddala, ou encore loi log-logistisque généralisée est une loi de probabilité continue dépendant de deux paramètres réels positifs c et k. Elle est communément utilisée pour étudier les revenus des ménages

Si X suit une loi de Burr (ou Sigh-Maddala), on notera $X \rightsquigarrow SM(c,k)$.

Caractérisation

La densité de peobabilité de la loi de Burr est donnée par :

$$f(x;c,k) = \begin{cases} ck\frac{x^{c-1}}{(1+x^c)^{k+1}} & si\ x > 0 \\ 0 & sinon \end{cases} \tag{1.39}$$

et sa fonction de répartition est :

$$F(x;c,k) = \begin{cases} 1-(1+x^c)^{-k} & si\ x > 0 \\ 0 & sinon \end{cases} \tag{1.40}$$

Si $c = 1$, la loi de *Burr* est la Distribution de *Pareto.*

Loi de Fréchet

La densité de probabilité de la distribution de paramètre $\alpha > 0$, peut être généralisée en introduisant un paramètre de position m du minimum et un paramètre d'échelle $s > 0$ prend la forme

$$f(x) = \begin{cases} \frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-1-\alpha} e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} & si\ x > m \\ 0 & sinon \end{cases}$$

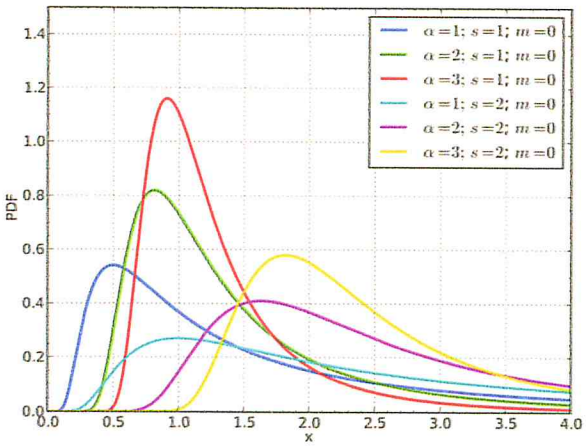


FIG. 1.9 – Densité de la Loi de Fréchet

Définition 1.11 *Sa fonction de répartition est est donnée par :*

$$\mathbb{P}(X \leq x) = \begin{cases} e^{-x^{-\alpha}} & si\ x > 0 \\ 0 & sinon \end{cases}$$

où $\alpha > 0$ est un paramètre de forme. Cette loi peut être généralisée en introduisant un paramètre de position m du minimum et un paramètre d'échelle $s > 0$. La fonction de répartition est alors :

$$\mathbb{P}(X \leq x) = \begin{cases} e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} & six > m \\ 0 & sinon \end{cases} \tag{1.41}$$

Propriétés :
Moments :
La loi de fréchet à un paramètre α a des moments standards :

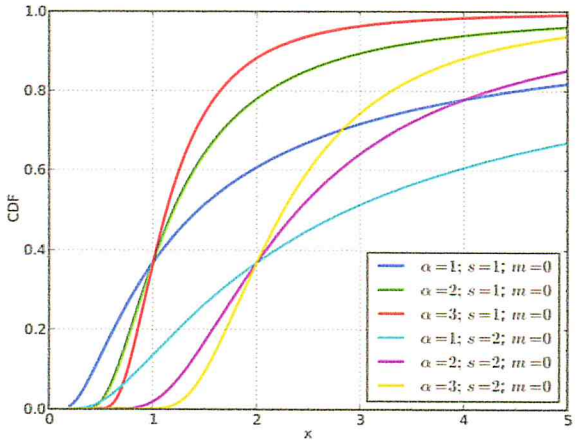


FIG. 1.10 – Fonction de répartition de la Loi de Fréchet

$$\mu_k = \int_0^\infty x^k f(x)\, dx = \int_0^\infty t^{-\frac{k}{\alpha}} e^{-t} dt \tag{1.42}$$

(avec $t = x^{-\alpha}$) définis pour $k < \alpha$:

$$\mu_k = \Gamma\left(1 - \frac{k}{\alpha}\right) \tag{1.43}$$

où $\Gamma(z)$ est la fonction Gamma.
En particulier :
– Pour $\alpha > 1$ l'**espérance** est

$$\mathbb{E}[X] = \Gamma\left(1 - \frac{1}{\alpha}\right) \tag{1.44}$$

– Pour $\alpha > 2$ la **variance** est

$$Var(X) = \Gamma\left(1 - \frac{2}{\alpha}\right) - (\Gamma\left(1 - \frac{1}{\alpha}\right))^2. \tag{1.45}$$

Quantiles Le **quantile** Q_Y d'ordre y peut être exprimé grâce à l'inverse de la fonction de répartition :

$$Q_Y = F^{-1}(y) = (-\log_e y)^{-\frac{1}{\alpha}}. \tag{1.46}$$

En particulier la **médiane** est :

$$Q_{1/2} = (\log_e 2)^{-\frac{1}{\alpha}}. \tag{1.47}$$

Le **mode** de la loi de *Fréchet* est :

$$\left(\frac{\alpha}{\alpha+1}\right)^{\frac{1}{\alpha}}. \tag{1.48}$$

Pour la loi de Fréchet à trois paramètres, le premier quartile est $Q_1 = m + \frac{s}{\sqrt[\alpha]{\log(4)}}$ est le troisième quartile est $Q_3 = m + \frac{s}{\sqrt[\alpha]{\lg(\frac{4}{3})}}$.

1.3 Théorèmes limites

Deux théorèmes mathématiques ont une place particulière en théorie des probabilités et en statistique : la loi des grands nombres et le théorème central limite

1.3.1 Loi (faible) des grands nombres

Théorème 1.2 *(LGN) Soient $X_1, ..., X_n$ des v.a.r. indépendantes, de même loi, et admettant une variance, on note $\mu = E(X_1)$. Alors, pour tout $\varepsilon > 0$*

$$p\left(\frac{X_1 + ... + X_n}{n} - \mu > \varepsilon\right) \overset{n \longrightarrow +\infty}{\longrightarrow} 0. \tag{1.49}$$

Dans ce cas, on dit que la moyenne arithmétique $\bar{X} = \frac{X_1+...+X_n}{n}$ converge en probabilité vers l'espérance mathématique lorsque n tend vers $+\infty$.

1.3.2 Théorème central limite (TCL)

Définition 1.12 *Soit $(Y_n)_{n\in N}$ une suite de v.a.r. et soit Y une v.a.r. On dit que $(Y_n)_{n\in N}$ converge en loi vers Y si pour tout x_0 point de continuité de la fonction de répartition F_Y de Y*

$$F_{Yn}(x_0) = p(Y_n \leq x_0) \underset{n\to+\infty}{\longrightarrow} F_Y(x_0) = p(Y \leq x_0). \tag{1.50}$$

On note la convergence en loi Y_n.

Théorème 1.3 *Soient $X_1, ..., X_n$ des v.a.r. indépendantes, de même loi, et admettant une variance, on note $\mu = E(X_1)$ et $\sigma = \sqrt{Var(X)}$. Alors :*

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \text{ lorsque } n \longrightarrow \infty.$$

1.4 Echantillonnages et Estimation

1.4.1 Echantillonnages

Un échantillon est un ensemble ordonné de données chiffrées rassemblées en vue de l'étude d'un phénomène, on note $(X_1, ..., X_n)$ cet ensemble. L'indice qui distingue les différentes valeurs est un entier qui court de 1 jusqu'à n la taille de l'échantillon, il représente l'ordre dans lequel les données sont considérées. D'un point de vue probabiliste, un échantillon est un n-uplet de variables aléatoires, ces variables peuvent être indépendantes ou dépendantes et être issues, ou non, de la même population parente, on dit qu'on a affaire à un échantillon de taille n ou à un n-échantillon. Parfois il est plus avantageux de considérer les X_i comme les n composantes d'un vecteur aléatoire X suivant une loi à n dimensions, dans ce dernier cas on interprête $(X_1, ..., X_n)$ comme un échantillon de taille 1 issu d'une population à n dimensions.

Les échantillons i.i.d.

L'échantillon pour lequel on possède le plus de résultats est l'échantillon « indépendant et identiquement réparti », où les X_i qui le composent sont des variables aléatoires mutuellement indépendantes et extraites de la même population. On dira de manière abrégée qu'un tel échantillon est i.i.d d'après l'anglais : "independent and identically distributed". On identifiera une série de résultats $(x_1, ..., x_n)$ obtenus indépendamment les uns des autres, et dans les mêmes conditions expérimentales, avec la réalisation d'un échantillon i.i.d de taille n.

A partir de maintenant et sauf mention expresse du contraire, il ne sera question que d'échantillons i.i.d.

Les échantillons ordonnés

On associe à un n-échantillon $(X_1, ..., X_n)$ un nouvel n-échantillon ordonné que l'on note $\left(\left(X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}\right)\right)$, et que l'on fabrique ainsi. On trie par ordre croissant les valeurs $(x_1, ..., x_n)$ des réalisations du n-échantillon.

Soit $\left(\left(x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}\right)\right)$ le résultat de ce tri. On considère alors les $x_{(i)}$ comme les réalisations d'une certaine variable aléatoire $X_{(i)}$.

Il est clair que les variables ordonnées $X_{(i)}$ ne suivent pas nécessairement la même loi que les X_i et que de plus elles ne sont pas indépendantes, même si au départ les X_i l'étaient. Dans un premier temps on précise la loi suivie par les $X_{(i)}$.

1.4.2 Estimation ponctuelle

On suppose que les données $x_1, ..., x_n$ sont n réalisations indépendantes d'une même variable aléatoire sous-jacente X. Il est équivalent de supposer que $x_1, ..., x_n$ sont les réalisations de variables al eatoires $X_1; ...; X_n$ indépendantes et de même loi. On notera θ le paramètre inconnu, et on notera $F(x; \theta)$ la fonction de répartition des X_i. Pour les variables aléatoires discrètes on notera $P(X = x; \theta)$ les probabilités élementaires, et pour les variables aléatoirescontinues on notera $f(x; \theta\ )$ la densité.

Définition d'un estimateur

Pour estimer θ on ne dispose que des données $x_1, .., x_n$, donc une estimation de θ sera une fonction de ces observations

Définition 1.13 *Une statistique t est une fonction des observations* $x_1, ...x_n$

$$\begin{array}{rcl} t & : & \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ (x_1, ..., x_n) & \mapsto & t(x_1, ..., x_n) \end{array} \tag{1.51}$$

Puisque les observations $x_1, ..., x_n$ sont des réalisations des variables aléatoires $X_1, ..., X_n$, la quantité calculable à partir des observations $t(x_1, ..., x_n)$ est une réalisation de la variable aléatoire $t(X_1, ..., X_n)$.Pour simplifer les écritures, on note souvent $t_n = t(x_1, ..., x_n)$ et $T_n = t(X_1, ..., X_n)$.

Définition 1.14 *Un estimateur d'une grandeur* θ *est une statistique* T_n *à valeurs dans l'ensemble des valeurs possibles de* θ *. Une estimation de* θ *est une réalisation* t_n *de l'estimateur* T_n.

Méthodes d'estimation

Nous ne nous intéressons qu'aux deux méthodes d'estimation les plus usuelles, la méthode des moments et la méthode du maximum de vraisemblance.

Mais il faut d'abord définir précisément ce que sont une estimation et surtout un estimateur.

L'estimateur des moments (EMM)

L'idée de base est d'estimer une espérance mathématique par une moyenne empirique,une variance par une variance empirique, etc...

Si le paramètre à estimer est l'espérance de la loi des X_i,alors on peut l'estimer par la moyenne empirique de l'échantillon.

Autrement dit, si $\theta = E(X)$, alors l'estimateur de θ par la méthode des moments (EMM) est

$$\bar{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Plus généralement, pour $\theta \in \mathbb{R}$, si $E(X) = \varphi(\theta)$ où φ est une fonction inversible, alors l'estimateur θ de par la méthode des moments est $\bar{\theta}_n = \varphi^{-1}(\bar{X}_n)$.

De la même manière, on estime la variance de la loi des X_i par la variance empirique de l'échantillon

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2.$$

Plus généralement, si la loi des X_i a deux paramètres θ_1 et θ_2 tels que

$$(E(X); Var(X)) = \varphi(\theta_1, \theta_2), \tag{1.52}$$

où φ est une fonction inversible, alors les estimateurs de θ_1 et θ_2 par la méthode des moments sont

$$(\bar{\theta}_{1n}, \bar{\theta}_{2n}) = \varphi^{-1}(\bar{X}_n - S_n^2). \tag{1.53}$$

Ce principe peut naturellement se généraliser aux moments de tout ordre, centrés ou non centrés : $E\left[X - E(X))^k\right]$ et $E(X^k), k \geq 1$.

La méthode du maximum de vraisemblance

La fonction de vraisemblance

Définition 1.15 *Quand les observations sont toutes discrètes ou toutes continues, on appelle fonction de vraisemblance (ou plus simplement vraisemblance) pour l' echantillon $x_1, ..., x_n$, la fonction du paramètre*

$$L(\theta, x_1, ..., x_n) = \begin{cases} P(X_1 = x_1, ..., X_n = x_n; \theta) & \text{si les } X_i \text{ sont discrètes} \\ f_{X_1,...,X_n}(x_1, .., ; x_n : \theta) & \text{si les } X_i \text{ dont continues} \end{cases} \tag{1.54}$$

Les X_i sont indépendantes et de même loi. Dans ce cas, la fonction de vraisemblance s'écrit

$$L(\theta, x_1, ..., x_n) = \begin{cases} \prod_{i=1}^n P(X_i = x_i, \theta) = \prod_{i=1}^n P(X = x_i, \theta) & \text{si les } X_i \text{ sont discrètes} \\ \prod_{i=1}^n f_{X_i}(x_i, \theta) = \prod_{i=1}^n f(x_i, \theta) & \text{si les } X_i \text{ dont continues} \end{cases} \tag{1.55}$$

L'estimateur de maximum de vraisemblance (EMV) En suivant le raisonnement précédent, pour n quelconque, il est logique de dire que la valeur la plus vraisemblable de θ est la valeur pour laquelle la probabilité d'observer $x_1, ..., x_n$ est la plus forte possible. Cela revient a faire comme si c'était l'éventualité la plus probable qui s'était produite au cours de l'expérience.

Définition 1.16 *L'estimation de maximum de vraisemblance de θ est la valeur $\hat{\theta}_n$ de qui θ rend maximale la fonction de vraisemblance $L(\theta; x_1, .., x_n)$. L'estimateur de maximum de vraisemblance (EMV) de θ est la variable aléatoire correspondante. Donc $\hat{\theta}_n$ sera en général calculé en maximisant la log-vraisemblance*

$$\hat{\theta}_n = \arg\max \ln \mathcal{L}(\theta; x_1, ..., x_n) \tag{1.56}$$

Quand $\theta = (\theta_1, ..., \theta_d) \in \mathbb{R}^d$ et que toutes les dérivées partielles ci-dessous existent, $\hat{\theta}_n$ est solution du système d'équations appelées équations de vraisemblance

$$\frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; x_1, ..., x_n) = 0 \qquad \forall j \in \{1, ..., d\}. \tag{1.57}$$

1.4.3 Estimateur à noyau

La construction d'un estimateur à noyaux

Rappelons que la densité de probabilité f est égale à la dérivée de la fonction de répartition F (si cette dérivée existe) . On peut donc écrire

$$\begin{aligned} f(x) &= \lim_{h \to 0} \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \lim_{h \to 0} \frac{P(x-h < X_i \leq x+h)}{2h} \end{aligned} \tag{1.58}$$

Un estimateur de $f(x)$ est alors :

$$\begin{aligned} \hat{f}(x) &= \frac{1}{2h} \frac{P(x-h < X_i \leq x+h)}{n} \\ &= \frac{1}{2hn} \sum_{i=1}^{n} I\{x-h < X_i \leq x+h\} \\ &= \frac{1}{2nh} \sum_{i=1}^{n} I\left\{-1 < \frac{X_i - x}{h} \leq +1\right\} \end{aligned} \tag{1.59}$$

Notons que cet estimateur peut encore s'écrire comme

$$\begin{aligned} \hat{f}(x) &= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}1_{\{x-h<X_i\leq x+h\}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K_0\left(\frac{X_i-x}{h}\right) \end{aligned} \tag{1.60}$$

Où

$$K_0(y) = \begin{cases} 1/2 & si\ y \in [-1,1[\\ 0 & sinon \end{cases} \tag{1.61}$$

Avec $K_0(.)$ La densité de probabilité uniforme sur l'intervalle $[-1,1[$ est appelée noyau de Rosenblatt[07] . Cet estimateur peut être généralisé en remplaçant la fonction de poids $K_0(.)$ par une fonction de poids plus générale K par exemple une densité de probabilité quelconque (Normale, Gamma, Bêta... etc)

Notion de noyau

Nous définitions maintenant plus généralement la notion d'estimateur à noyau

Définition 1.17 *Soit $K : \mathbb{R} \to \mathbb{R}^+$ une fonction intégrable telle que $\int_R K(u)\,du = 1$, K est dit noyau, pour $n \in \mathbb{N}^*$, on appelle $h_n > 0$ la fenêtre ou paramétré de lissage et $\hat{f}(x)$ l'estimateur à noyau de la densité de probabilité f définit pour tout $x \in \mathbb{R}$ par*

$$\hat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h_n}\right) \tag{1.62}$$

On note $h_n = h$.

Définition 1.18 *Un noyau est dit symétrique si, pour tout u dans son ensemble de définition $K(u) = K(-u)$,$\forall u$ dans sa domaine de définition. ce qui implique l'égalité suivante :*

$$\int_{\mathbb{R}} uK(u)\,du = 0 \tag{1.63}$$

De plus,elle est de carré intégrable

$$\int_{\mathbb{R}} K^2(u)\,du < +\infty \tag{1.64}$$

et nous avons aussi la variance de K finie

$$\int_{\mathbb{R}} u^2 K(u)\, du < +\infty \tag{1.65}$$

Exemple 1.4 *Noyau gaussien*

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), \qquad u \in \mathbb{R}.$$

Chapitre 2: Indices des Inégalités

Chapitre 2

Indices des Inégalités

2.1 Introduction

Au cours des dernières décennies, les inégalités ont joué un rôle important dans de nombreuses branches des sciences sociales, principalement la sociologie et l'économie, étant l'une des clés les enjeux du discours sur le bien-être des sociétés et des individus . Ainsi, il apparaît une question de comment mesurer ces inégalités de manière appropriée. Il existe de nombreux indices d'inégalité dans la littérature le plus populaire, à savoir, l'indice de Gini, l'indice de Theil et la mesure d'Atkinson . De ceux-ci, l'indice de Gini est le plus souvent utilisé et également mieux connu des non-scientifiques. Récemment, un nouvel indice d'inégalité a été proposé par Zenga. Il a toutes les propriétés qui sont généralement requises pour les mesures d'inégalité. Afin de décider quelle mesure d'inégalité est la plus appropriée pour un sujet donné, il serait utile d'étudier et de comparer les propriétés de différents indices.

2.2 Définition de revenu

Etymologie : de revenu, composé du préfixe, indiquant un retour à un état initial et du latin venire, aller, venir, arriver.

En économie, un revenu est l'ensemble des ressources ou droits qu'un individu, une entreprise ou une collectivité publique, perçoit sur une période donnée, en nature ou en monnaie, sans prélever sur son patrimoine.

Synonymes : allocation, gain, pension, produit, rente, rétribution, salaire.

Contrairement au patrimoine qui est un stock de biens détenus à un instant donné, le revenu est un flux de biens et services dont on dispose pendant une période donnée. En outre, pour parler de revenus récurrents, par

opposition aux revenus exceptionnels, ceux-ci doivent se répéter de période en période.

On distingue les sommes perçues au titre de :

- la rémunération pour un travail. Ex : salaire.
- du patrimoine. Ex : loyers d'immeubles, produit d'un capital placé (intérêts, dividendes, redevance d'utilisition de brevet).
- l'activité : Services rendus et produits fournis par les professionnels et entreprises.
- des prestations et transferts sociaux. Ex : indememnités de chômage, allocations sociales...

"Le revenu salarial correspond à la somme de tous les salaires perçus par un individu au cours d'une année donnée, nets de toutes cotisations sociales, y compris Contribution Sociale Généralisée (CSG) et contribution au remboursement de la dette sociale (CRDS)".

Le **revenu net** est le revenu brut diminué des dépenses occasionnées pour sa perception (frais professionnels, entretretien d'un patrimoine, etc...).

Le **revenu disponible** d'un ménage est l'ensemble de ses revenus d'activité, de son patrimoine, et des prestations et transferts sociaux perçus, nets des impôts directs (impôt sur le revenu, taxe d'habitation, CSG, CRDS).

Le **revenu réel** correspond au pouvoir d'achat réel, c'est-à-dire en tenant compte des variations des prix des biens et des services.

Le **revenu national brut** (RNB) est la somme des revenus perçus, pendant une période donnée, par les agents économiques résidant sur le territoire national. Il est la somme du PIB et du solde des flux derevenus primaires avec le reste du monde.

Le **revenu par tête** (ou RNB par habitant) est le revenu national brut (RNB) annuel, divisé par le nombre total d'habitants, pour un pays ou une région donnée.

2.3 Courbe de Lorenz

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 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.

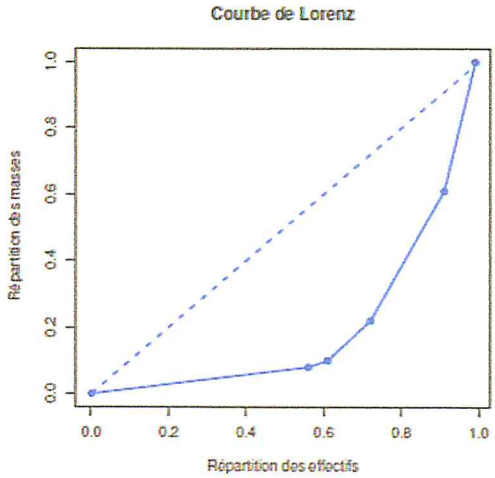


FIG. 2.1 – Courbe de Lorenz

Algèbriquement, on a la relation suivante pour la fréquence cumulée des effectifs :

$$p_i = \frac{1}{N}\sum_{j=1}^{i} n_j = \frac{1}{N}(n_1 + n_2 + ... + n_i) \tag{2.1}$$

avec $N = n_1 + n_2 + ... + n_k$.

De même, on a la relation suivante pour la fréquence cumulée des masses $n_i v_i$:

$$q_i = \frac{1}{T}\sum_{j=1}^{i} n_j v_j = \frac{1}{T}(n_1 v_1 + n_2 v_2 + ... + n_i v_i) \tag{2.2}$$

avec $T = n_1 v_1 + n_2 v_2 + ... + n_k v_k$.

Par convention, on pose $p_0 = q_0 = 0$.

Un point de coordonnées (p, q) sur la courbe de Lorenz indique que $p\%$ des individus se partagent $q\%$ de la masse.

La bissectrice du carré est la ligne d'équirépartition. C'est ce que serait la courbe de concentration s'il y avait équirépartition des masses. Sur cette diagonale, en tout point, $p\%$ des individus se partageraient exactement $p\%$ de la masse. Dans ce cas, la concentration est nulle.

Définition 2.1 *L'aire de concentration est la région comprise entre le diagonale et la courbe de lorenz.*

Interprétation : plus cette aire est importante, c'est-à-dire plus la coube de concentration s'écarte de la bissectrice, plus la concentration est forte

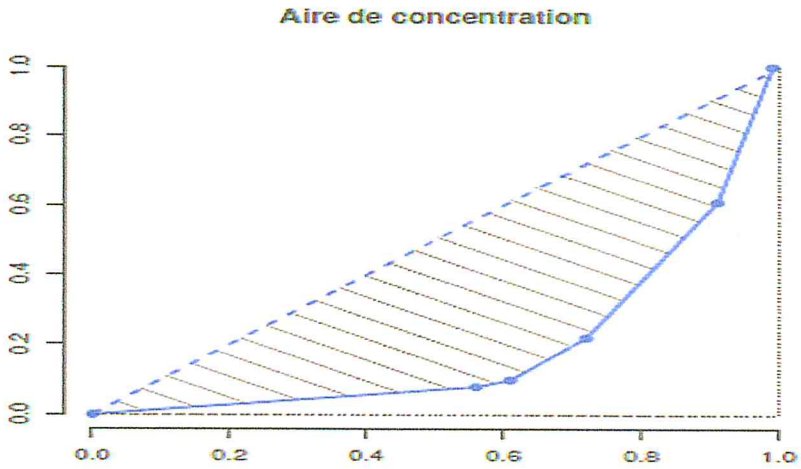


FIG. 2.2 – Aire de concentration

$$L(p) = \frac{1}{\mu} \int_0^p Q(q)\, dq \tag{2.3}$$

avec $Q(q)$ représentant la répartition des revenus, μ le revenu moyen et pour des valeurs de p variant de 0 à 1. Lorsque $L(0,5) = 0,3$ on en déduira que 50 % des individus les plus modestes possèdent 30 % du revenu total.

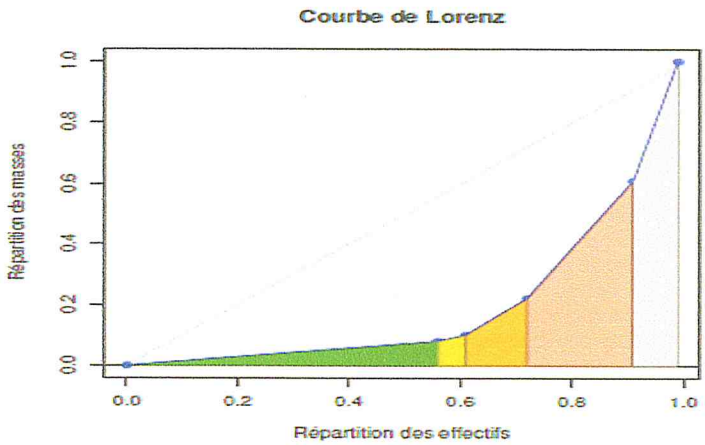


FIG. 2.3 – Courbe de Lorenz

2.4 Indice de Gini

L'indice de Gini a été élaboré par Gini en 1912 et entretient un lien strict avec la représentation de l'inégalité des revenus à l'aide de la courbe de Lorenz. En particulier, il mesure le ratio entre l'aire située entre la courbe de Lorenz et la droite d'équidistribution (et donc l'aire de concentration) et l'aire de concentration maximale.

La figure (2.4) représente ces aires : elle trace trois courbes de Lorenz à partir de trois distributions de revenus hypothétiques O, P et Q

La courbe basée sur la distribution des revenus O est la courbe standard que donne l'analyse des distributions de revenus réelles.

Celle de la distribution P représente le cas extrême où tous les revenus sont égaux. Dans ce cas, elle prend aussi le nom de droite d'équidistribution.

Enfin, la courbe de la distribution Q illustre un autre cas extrême, celui où tous les revenus sont nuls, sauf le dernier.

Dans la (2.4), OP est la droite d'équidistribution et ORP l'aire définie par la courbe de Lorenz de la distribution des revenus standard et la courbe d'équidistribution, baptisée aire de concentration. OPQ est l'aire de concentration maximale, c'est-à-dire la zone entre la courbe de Lorenz de la distribution de revenus Q et la droite d'équidistribution.

La droite d'équidistribution OP et l'aire OPQ représentent les valeurs extrêmes de l'aire de concentration dans une courbe de Lorenz. Soit cette aire est nulle (comme dans le cas de la droite d'équidistribution de la distribution P), soit elle est maximale (cas de la distribution Q). Pour une distribution des revenus standard, l'aire de concentration se situe quelque part entre zéro et l'aire de concentration maximale, comme dans la figure (2.4).

L'indice de Gini mesure le ratio entre l'aire de concentration et l'aire de concentration maximale. Par conséquent, dans la figure (2.4) :

$$G = \frac{aire\ de\ concentration}{aire\ de\ concentration\ maximale} = \frac{ORP}{OPQ} \tag{2.4}$$

Comme l'aire de concentration maximale correspond à une distribution où un seul individu détient la totalité des revenus, l'indice de Gini G mesure en général la distance entre l'aire définie par une quelconque distribution de revenus standard et l'aire de concentration maximale.

Il faut maintenant comprendre comment s'applique la formule de la figure (2.4) dans la pratique. Commençons par le dénominateur de G. Que les coordonnées maximales de la courbe de Lorenz se situent au point (1,1). Par conséquent, l'aire OPQ doit être un triangle possédant une longueur de base

de 1 et une hauteur de 1. Son aire est donc égale à $\frac{1}{2}$. Le dénominateur de G est donc $\frac{1}{2}$.

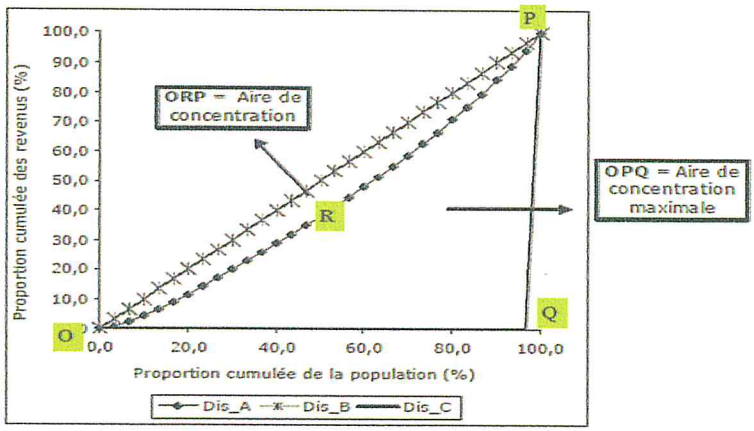


FIG. 2.4 – Courbe de Lorenz et indice de Gini

$$GINI = \frac{Aire\ de\ concentration}{Aire\ de\ concentration\ maximale} = \frac{ORP}{OPQ} \tag{2.5}$$

Mais qu'en est-il du numérateur ? Au lieu de calculer directement l'aire de concentration, nous pouvons exploiter le fait que cette aire représente la différence entre l'aire de concentration maximale et l'aire sous la courbe de Lorenz (cette dernière étant donnée par ORPQ). Le mode de calcul le plus facile de l'aire sous la courbe de Lorenz est décrit ci-après.

Commençons par rappeler la définition des coordonnées de la courbe de Lorenz. Si $y_1 \leq y_2 \leq ... \leq y_n$:

$$q_i = \frac{y_1 + y_2 + ... + y_i}{y_1 + y_2 + ... + y_n} = \frac{y_1 + y_2 + ... + y_i}{Y} \rightarrow proportion\ cumulée\ des\ revenus \tag{2.6}$$

$$p_i = \frac{i}{n} \rightarrow proportion\ cumulée\ de\ la\ population \tag{2.7}$$

où $q_0 = p_0 = 0$ et $q_n = p_n = 1$.

L'aire sous la courbe de Lorenz ORPQ est la somme des aires d'une série de polygones. Regardons la figure (2.5), où une courbe de Lorenz simplifiée a été crée pour une population de quatre individus. Le premier polygone est un triangle (PQO) et les trois autres sont des trapèzes isocèles pivotés. On

peut donc calculer chaque aire séparément et ajouter les résultats obtenus pour obtenir la valeur de l'aire globale. Symbolisons l'aire du ième polygone par Z_i et l'aire totale obtenue de cette manière par Z.

L'aire du triangle est donnée par :

$$Z_1 = \frac{\overbrace{p_1}^{base}\ \overbrace{q_1}^{hauteur}}{2} \tag{2.8}$$

tandis que l'aire de chaque trapèze est donnée par :

$$Z_i = \frac{\overbrace{(q_i + q_{i-1})(\overline{p_i} - \overline{p_{i-1}})}^{\text{(base longue + base coute) } \times \text{hauteur}}}{2} \tag{2.9}$$

Comme $q_0 = p_0 = 0$, la somme de toutes ces aires donne :

$$Z = \sum_{i=1}^{n} Z_i = \frac{1}{2}\sum_i [(q_i + q_{i-1})(p_i - p_{i-1})] \quad pour\ n = 4 \tag{2.10}$$

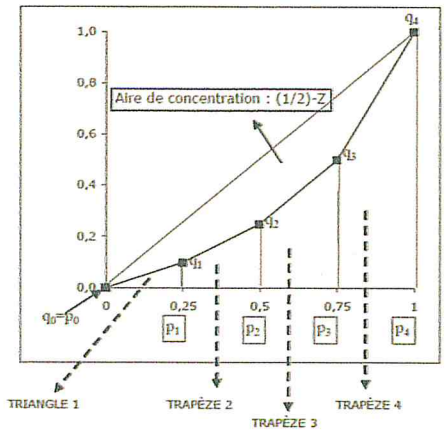


FIG. 2.5 – Mode de calcul de l'aire de concentration

Cependant, Z n'est pas l'aire de concentration, mais l'aire sous la courbe de Lorenz. Pour calculer l'aire de concentration (numérateur de l'indice de Gini), il suffit maintenant de soustraire Z de l'aire de concentration maximale ($\frac{1}{2}$) comme suit :

$$Aire\ de\ concentration = \frac{1}{2} - Z = \frac{1}{2} - \frac{1}{2}\sum_i [(q_i + q_{i-1})(p_i - p_{i-1})] \tag{2.11}$$

Selon (4), l'indice de Gini G est donc égal à :

$$G = \frac{\frac{1}{2} - \frac{1}{2}\sum_i [(q_i + q_{i-1})(p_i - p_{i-1})]}{\frac{1}{2}} = 1 - \sum_i [(q_i + q_{i-1})(p_i - p_{i-1})] \tag{2.12}$$

que l'on peut également écrire :

$$G = 1 - 2Z. \tag{2.13}$$

La formule ci-dessus indique seulement que l'indice de Gini est égal à 1 moins deux fois l'aire sous la courbe de Lorenz.

Cette interprétation géométrique basée sur la courbe de Lorenz ne constitue que l'un des modes de calcul possibles de l'indice de Gini. Une autre approche, qui va s'avérer particulièrement utile ci-après, consiste à exprimer directement l'indice de Gini en termes de covariance entre les niveaux de revenus et la distribution cumulée des revenus.

En particulier :

$$G = Cov(y, F(y))\frac{2}{y} \tag{2.14}$$

où Cov représente la covariance entre des niveaux de revenus y et la distribution cumulée des mêmes revenus $F(y)$ et où $\overline{y}$ est le revenu moyen. Il est utile de rappeler ici que la covariance est la valeur attendue E des produits des écarts sur la moyenne de chaque variable. Soit dans ce cas précis :

$$Cov\,[y, F(y)] = E([y - \overline{y}] \cdot \left| F(y) - \overline{F(y)} \right|) \tag{2.15}$$

En général, l'index de Gini est une fonction $G : R_n^+ \to [0, 1]$ qui attribue à chaque vecteur de revenu non négatif un nombre réel compris entre 0 et 1, ce qui représente le niveau d'inégalité de la société. Cette mesure est 0 en égalité maximale et 1 en parfaite inégalité. La dénotation attrayante de l'indice de Gini est le double de la surface entre la ligne d'égalité et la courbe de Lorenz dans la boîte de l'unité. La ligne à 45° représente l'égalité parfaite des revenus et la zone située entre cette ligne et la courbe de Lorenz est appelée zone de concentration. Par conséquent, l'indice de Gini peut être exprimé comme

$$G = 2\int_0^1 (p - L(p))\, dp, \tag{2.16}$$

tel que $p = F(x)$ est une fonction de distribution cumulative (fdc) non-négatif revenu avec espérance positive et négative μ, $L(p)$ la fonction de

Lorenz donnée par

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)\,dt, \quad où\ F^{-1}(t) = \inf\{x \mid F(x) \geq p : p \in [0,1]\}. \tag{2.17}$$

En utilisant la définition de l'indice de Gini dans l'équation (2.16), soit le double de la surface entre la droite d'égalité et la courbe de Lorenz, et en appliquant un changement de variable $p = F(x)$, on peut constater que :

$$G = \frac{2}{\mu}\int_0^\infty xF(x)\,dF(x) - 1. \tag{2.18}$$

Supposons qu'un échantillon i.i.d de taille n est tiré au hasard de la population, et $\hat{F}$ désigne la fonction de distribution empirique correspondante. Soit $X_1, ..., X_n$ être un échantillon aléatoire et $X_{1:n} \leq ... \leq X_{n:n}$ les statistiques d'ordre obtenues à partir de l'échantillon. En suite, un estimateur alternatif du coeffcient de Gini peut être obtenu par le fdc empirique $\left(\hat{F}\right)$ du revenu au lieu de sa fonction distribution correspondante F dans (2.18) comme :

$$\hat{G} = \frac{2}{\hat{\mu}}\int_0^\infty x\hat{F}(x)\,d\hat{F}(x) - 1 \tag{2.19}$$

à cet égard, l'exemple de l'indice de Gini peut être exprimé comme

$$\begin{aligned}\tilde{G} &= \frac{2}{\hat{\mu}}\int_0^\infty xd\hat{F}(x))^2 - 1, \\ &= \frac{2\sum_{i=1}^n X_{i:n}\left(i - \frac{1}{2}\right)}{n\sum_{i=1}^n X_i} - 1.\end{aligned} \tag{2.20}$$

Davidson (2009) a trouvé une expression approchée du biais de $\hat{G}$ à partir du quel il a dérivé l'estimateur à correction de biais du coefficient de Gini noté $\tilde{G}$; lequel est donné par :

$$\tilde{G} = \frac{n}{n-1}\hat{G}, \tag{2.21}$$

alors que l'estimateur (2.21) est toujours biaisé mais son biais est d'ordre n^{-1}, il est parfois recommandé d'utiliser cet estimateur parceque l'estimateur correctement corrigé du biais est non seulement plus facile à calculer que les autres estimateurs mais aussi son biais converge vers 0 plus vite que $n \to \infty$.

L'indice de Gini généralisé a été popularisé par les travaux de Donaldson et Weymark (1980), et de Yitzhaki (1983) :

$$I_\rho = \frac{\mu - \zeta_\rho}{\mu} \tag{2.22}$$

avec

$$\zeta_\rho = \sum_{i=1}^{L} \left[\frac{(R_i)^\rho - (R_{i+1})^\rho}{(R_i)^\rho} \right] y_i \tag{2.23}$$

et

$$R_i = \sum_{i=1}^{L} \omega_i \tag{2.24}$$

Où μ représente la moyenne des revenus, ω et y_i le poids et le niveau de revenu de l'individu, et ρ le paramètre d'aversion à l'inégalité.

2.5 Indicateur de Theil

L'indice de Theil (1967) mesure l'écart entre le poids d'un individu (ou d'un groupe) dans la population et le poids de son revenu dans le revenu total . L'indice de Theil repose sur le concept physique d'entropie [Figini, 1998]. Cet indice correspond à la variation d'entropie entre la situation parfaitement égalitaire et la situation réelle. « En thermodynamique, l'entropie définit l'état de désordre d'un système, croissant lorsque celui-ci évolue vers un état de désordre accru » (Petit Robert, 1986). Sa valeur varie entre 0, la situation d'égalité et $logN$, dans le cas où tous les revenus sont nuls, sauf un.
Soit y_i le revenu de l'individu i appartenant à une population de N individus et μ le revenu moyen, l'indice s'écrit :

$$T = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{\mu} (\log(\frac{y_i}{\mu})) \tag{2.25}$$

Cet indice accorde un peu plus d'importance à l'inégalité dans le bas de la distribution qu'à l'inégalité parmi les riches. Moins couramment utilisé que l'indice de Gini, l'indice de Theil présente néanmoins des atouts pratiques incontestables. Son principal intérêt est de pouvoir se décomposer à l'infini en partitionnant la population puis en redécomposant chacun des groupes en différents sous-groupes, cela afin d'analyser l'évolution des inégalités dans et entre différentes sous-populations. Cependant son expression mathématique, qui utilise la forme logarithmique, limite son usage à des valeurs non nulles.

2.6 Indice d'Atkinson (1970)

L'indice d'Atkinson est un indice de l'inégalité des revenus basé sur la théorie économique. Afin de définir son indice d'inégalité, Atkinson (1970) suppose que le bien-être dans la société peut être évalué à partir de l'équation suivante :

$$W = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \frac{x_i^{1-\varepsilon}}{1-\varepsilon} & pour\ \varepsilon \neq 1 \\ \frac{1}{n}\sum_{i=1}^{n} \ln x_i & pour\ \varepsilon = 1 \end{cases} \tag{2.26}$$

Le paramètre ε décrit l'aversion de la société pour l'inégalité.

- Si $\varepsilon = 0$ il n'y a aucune aversion à l'inégalité. Qu'il soit distribué à un riche ou à un pauvre, un euro supplémentaire augmente pareillement le bien-être social W.
- Si $\varepsilon \to \infty$ l'aversion à l'inégalité est extrême et le bien-être social est confondu avec le bien-être de l'individu le plus pauvre (Rawls)

Les indices d'inégalité associés à la fonction de bien-être social d'Atkinson s'écrivent :

$$I_A = \begin{cases} 1 - \left(\frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i}{\mu} \right)^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}} & si \quad \varepsilon \neq 1 \\ 1 - \prod_{i=1}^{n} \left(\frac{x_i}{\mu} \right)^{\frac{1}{n}} & si \quad \varepsilon = 1 \end{cases} \tag{2.27}$$

où μ représente la moyenne des revenus observés. A l'aide de l'utilisation de la moyenne généralisée

$$M^{\varepsilon} = \begin{cases} \sqrt[\varepsilon]{\frac{1}{n}\sum_{i=1}^{n} x_i^{\varepsilon}} & si\ \varepsilon \neq 0 \\ (\Pi_{i=1}^{n} x_i) & si\ \varepsilon = 0, \end{cases} \tag{2.28}$$

l'indice d'Atkinson peut s'exprimer par :

$$A_{\varepsilon} = \frac{M^1 - M^{1-\varepsilon}}{M^1} = 1 - \frac{M^{1-\varepsilon}}{M^1}. \tag{2.29}$$

L'indice d'Atkinson est donc fonction du paramètre ε. Pour cette étude, nous avons utilisé l'indice

pour deux valeurs du paramètre : $\varepsilon = 0.5$ et $\varepsilon = 1$:

$$A_{1/2} = 1 - \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i}\right)^2}{\mu}, \tag{2.30}$$

$$A_1 = 1 - \prod_{i=1}^{n}\left(\frac{x_i}{\mu}\right)^{\frac{1}{n}} = 1 - \frac{G}{\mu},$$

avec G la moyenne géometrique :

$$G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}. \tag{2.31}$$

En pratique on interprète le coefficient ε en remarquant que plus ce paramètre décroît, plus on attache d'importance aux transferts concernent les revenus les plus faibles.

2.7 Indice Bonferroni

C.E. Bonferroni (1930) a proposé une mesure de l'inégalité des revenus, basée sur des moyens partiels, ce qui est souhaitable lorsque la principale source d'inégalité de revenu est la présence d'unités dont le revenu est très inférieur à celui des autres.

Le but est de passer en revue les propriétés théoriques et statistiques de l'indice de Bonferroni et de les relier aux caractéristiques de la distribution des revenus. Les principaux résultats obtenus peuvent être résumés comme suit :

a) L'indice se concentre sur les bas revenus.

b) L'indice satisfait au principe de transfert décroissant introduit par Mehran (1976).

c) L'indice n'est pas additionnellement décomposable.

Définition de l'indice de Bonferroni pour les distributions continues.

Soit Y une variable aléatoire continue non négative avec fonction de distribution cumulative $F(Y)$. La moyenne partielle (ou conditionnelle) de Y sur l'intervalle $[0, y]$ est donnée par

$$m(y) = \frac{\int_0^y u dF(u)}{F(y)} \tag{2.32}$$

Pour un niveau donné y du revenu Y

$$r(y) = \frac{\mu - m(y)}{\mu};\ 0 < \mu < \infty \tag{2.33}$$

est une fonction bornée, monotone décroissante et non négative dans $[0, \infty[$et mesure la différence relative entre le revenu moyen total μ et la moyenne des revenus inférieurs ou égale à y. La moyenne (2.33) de tous les revenus donne l'indice de Bonferroni

$$B = \int_0^{\infty} r(y)\, dF(y) \tag{2.34}$$

Notez que comme suggéré Pizzetti (1955) l'indice de Gini R peut être exprimé comme

$$R = \int_0^{\infty} r(y) \left[\frac{F(y)}{\int_0^{\infty} F(y)\, dF(y)} \right] dF(Y) \tag{2.35}$$

par conséquent R est la moyenne pondérée des $r(Y)$ alors que B est leur moyenne simple. Puisque $r'(Y)$ et $F'(Y)$ ont un signe opposé, alors $B \geq R$ (voir De Vergottini, 1940). La moyenne progressive $m(Y)$ est égale au rapport $\mu F_1(Y)/F(Y)$, où $F_1(Y)$ est la distribution incomplète de premier instant correspondant à F. Par conséquent, la formule (??) peut aussi être écrit comme

$$B = \int_0^{\infty} \left[\frac{F(y) - F_1(y)}{F(y)} \right] dF(y) = 1 - \int_0^{\infty} \left[\frac{F_1(y)}{F(y)} \right] dF(y) = 1 - \int_0^{\infty} F_1(y)\, d\ln[F(y)] \tag{2.36}$$

L'indice de Bonferroni peut également être considéré comme la statistique récapitulative de la courbe de Bonferroni de la distribution des revenus. Une telle courbe notée $B(p) : [0,1] (\supset R) \rightarrow [0,1]$, est définie (voir Zenga, 1984a) comme la relation entre la proportion cumulée $p = F(y)$ des unités de revenu (IRU) et le ratio de la part cumulée du revenu $q = F(y)$ et p . Soit :

$$F^{-1}(t) = \inf\{y : F(y) > t\} \tag{2.37}$$

est la fonction inverse de $F(y)$ et $F(y)$ respectivement.

alors :

$$B(p) = \int_0^{p} \left[\frac{F_1^{-1}(t)}{F^{-1}(t)} \right] dt \tag{2.38}$$

La courbe de Bonferroni est représentée dans un carré unitaire. Il est facile de vérifier que$B(0) = 0$, $B(1) = 1$, et que

$B(p)$ est une fonction non décroissante pour $p \in [0,1]$. Clairement l'égalité parfaite aboutirait à des points le long de la ligne $B(p) = 1$, et si une IRU avait tous les revenus, la courbe de Bonferroni coïnciderait avec les cathètes OW et WZ.

D'un point de vue géométrique, l'indice de Bonferroni est la zone située entre la courbe de Bonferroni et la ligne d'égalité parfaite

$$B = 1 - \int_0^1 B(p)\, dp \tag{2.39}$$

Définition de l'indice de Bonferroni pour les distributions discrètes.

La population est supposée être composée de n IRU qui sont étiquetées dans l'ordre non-croissant du revenu de sorte que l'indice i indique le rang de y_i parmi $y_1, y_2, ..., y_n$. Soit μ le revenu moyen arithmétique, P_i soit la part cumulée de la population et Q_i la part cumulée des revenus correspondant aux premières IRU. Ainsi

$$P_i = \frac{i}{n}; Q_i = \frac{1}{n\mu}\sum_{j=1}^{i} y_j \quad (i = 1, 2, ..., n) \tag{2.40}$$

La moyenne des revenus inférieurs ou égaux à y_j est $M_i = \frac{1}{i}\sum_{j=1}^{i} y_j$; par conséquent pour un discrèt distribution, nous avons

$$B_i = \frac{1}{n-1}\sum_{i=1}^{n-1}\left[\frac{\mu - M_i}{\mu}\right] = \frac{1}{n-1}\sum_{i=1}^{n-1}\left[\frac{P_i - Q_i}{P_i}\right] \tag{2.41}$$

ces expressions montrent que B est facilement estimable à partir de sources de données existantes. Alternativement B_n peut être écrit comme un rapport de combinaisons linéaires de statistiques d'ordre

$$B_n = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} y_i} \tag{2.42}$$

avec

$$w_i = 1 - \sum_{j=1}^{n}\frac{1}{j};\ w_{i+1} = w_i + \frac{1}{i}; \sum_{i=1}^{n} w_i = 0 \tag{2.43}$$

Cette spécification aide à caractériser le système de pondération des revenus dans la fonction de bien-être derrière l'indice de Bonferroni. Si un

transfert d'une unité de revenu préservant le rang a lieu de la s-ème à la r-ème IRU (avec $s > r$) B_n changera d'un montant

$$\Delta B_n = -\frac{1}{(n-1)\,\mu}\sum_{j=r}^{s-1}\frac{1}{j} \tag{2.44}$$

qui est proportionnelle au nombre d'IRU dont le revenu tombe en $[y_r, y_{s-1}]$. En outre la formule (2.44) montre que pour une différence fixe $(s-r)$ entre les deux rangs, le plus bas est r le plus haut est ΔB_n pour que le Bonferroni satisfasse le principe de transfert décroissant : un petit transfert positif d'une unité plus riche vers une unité plus pauvre diminue l'inégalité et la diminution est d'autant plus importante que l'unité est pauvre. Il faut noter, cependant, que l'effet du transfert dépend uniquement des rangs r et s et non de la taille des niveaux de revenu (voir aussi Salvaterra, 1986).

De Vergottini (1940) a interprété (2.42) comme la fraction du revenu total qui devrait être transférée pour atteindre un état d'égalité parfaite au moyen d'un nivellement graduel des revenus à partir de l'IRU la plus pauvre. En fait, pour égaler y_1 et y_2, la quantité M_2 doit être soustraite de y_2 et $(M_2 - M_1)$ ajoutée à y_1 . De même, pour éliminer les différences entre $y_1 = M_2, y_2 = M_2$ et y_3, la quantité M_3 doit être soustraite de y_3 et $(M_3 - M_2)$ ajoutée à chaque M_2. En général, $i\,(M_{i+1} - M_i)$ est transféré à la i-ème redistribution de sorte que le revenu transféré pendant tout le processus de péréquation est

$$\sum_{i=1}^{n-1} i\,(M_{i+1} - M_i) = \sum_{i=1}^{n-1}(\mu - M_i) \tag{2.45}$$

en divisant (2.45) par le revenu qui doit être transféré dans le cas d'une inégalité complète, c'est-à-dire $(n-1)\,\mu$, on obtient B_n.

Il est facile de vérifier que l'indice de Bonferroni vérifie les propriétés suivantes :

1. $0 \leq B \leq 1$.
2. $B = 0$ si et seulement si tous les revenus sont égaux.
3. $B = 1$ si et seulement si un seul revenu est positif.
4. B est indépendant de l'échelle.
5. B est symétrique (ne dépend pas de l'affectation des étiquettes à l'IRU)
6. Les additions égales (soustractions) diminuent (augmentent) B
7. B ne satisfait pas la propriété d'invariance de la réplication de la population (Zenga, 1986) a prouvé qu'il existe une incompatibilité logique entre la troisième propriété et la propriété d'invariance à la réplication de la population)

8. B appartient à la classe des mesures linéaires de l'inégalité des revenus

$$J = \frac{\int_0^1 \left[F^{-1}(p) - \mu\right] W(p)\, dp}{\mu}, \int_0^1 W(p)\, dp = 0$$

défini par Mehran (1976) avec $W(p) = 1 + Log(p)$. Mehran (1976) a examiné la fonction de score linéaire $2(p-1)$ correspondant à l'indice de Gini et la fonction de score quadratique $W(p) = 1 - 3(1-p)$ qui ne correspond pas à une mesure d'inégalité bien connue. "Peu d'autres choix de $W(p)$ semblent avoir été explorés" (Arnold, 1983). Les deux auteurs ont omis de mentionner l'indice de Bonferroni.

2.8 Indice d'inégalité de Zenga

Zenga propose une mesure d'inégalité basée sur la courbe d'inégalité $I(p)$, définie en termes de moyennes arithmétiques inférieure et supérieure d'une distribution. Cette idée de mesure de l'inégalité des revenus consiste à comparer les moyennes arithmétiques des revenus de deux groupes, appelés groupes inférieurs et supérieurs. La division des données ordonnées en deux groupes est faite en choisissant un point de division. un extrême, le groupe inférieur consiste seulement en l'observation la plus basse. l'autre extrême, le groupe supérieur ne comprend que les revenus les plus élevés. Soit

$$\left\{(x_j, n_j) : j = 1, ..., s; \quad 0 \leq x_1 \leq x_2 \leq ... \leq x_s; \quad \sum_{j=1}^{s} n_j = N\right\} \tag{2.46}$$

dénote la distribution de fréquence d'une variable aléatoire non négative X. En suite, divisons cette distribution en deux parties, respectivement le groupe inférieur et le groupe supérieur :

$$\{(x_1, n_1), (x_2, n_2), ..., (x_j, n_j)\}, \{(x_j, n_j), (x_{j+1}, n_{j+1}), .., (x_s, n_s)\}. \tag{2.47}$$

Pour chaque point de division, il est possible de définir la moyenne inférieure $\overline{M}(p_j)$et la moyenne supérieure $M^+(p_j)$ de la distribution divisée comme suit :

$$\begin{aligned} \overline{M}(p_j) &= \frac{1}{N_j}\sum_{i=1}^{j} x_i n_i, \ j = 1, ..., s \qquad (2.48) \\ M^+(p_j) &= \frac{1}{N - N_{j-1}}\sum_{i=j}^{s} x_i n_i, \ j = 1, ..., s \end{aligned}$$

où $N_j = \sum_{i=1}^{j} n_i$ et $p_j = \frac{N_j}{N}$.

Comparons $\overline{M}(p_j)$ et $M^+(p_j)$ en utilisant l'indice, nous obtenons une mesure ponctuelle de l'uniformité de la distribution. $(U(p_j)) \times 100$ donne la moyenne inférieure en pourcentage de la moyenne supérieure. L'indice d'inégalité de point est défini en terme de $U(p_j)$ comme :

$$I(p_j) = 1 - U(p_j) \tag{2.49}$$

La mesure d'inégalité synthétique proposée par Zenga est la moyenne arithmétique pondérée suivante du point mesures $I(p_j)$:

$$Z = \sum_{j=1}^{s} I(p_j) \frac{n_j}{N} \tag{2.50}$$

$U(p_j)$ et $I(p_j)$ prennent des valeurs comprises entre 0 et 1 inclusivement. En particulier :

$$U(p_1) = \frac{x_1}{M}, \quad U(p_s) = \frac{M}{x_s}, \quad I(p_1) = 1 - \frac{x_1}{M}, \quad I(p_s) = 1 - \frac{M}{x_s} \tag{2.51}$$

où M est la moyenne de toutes les observations.
L'indice de Zenga prend la valeur 0 dans le cas d'aucune inégalité.
La forme de la courbe $I(p_j)$ en fonction de p_j n'est pas contrainte par des points fixes $(0,0)$ et $(1,1)$, comme dans le cas de la courbe de Lorenz .

- **Indice de Zenga cas continu**

Dans le cas continu, l'indice de Zenga est donné par

$$Z = 1 - \int_0^1 \frac{L(\alpha)}{\alpha} . \frac{1-\alpha}{1-L(\alpha)} d\alpha. \tag{2.52}$$

- Nouvel mesure d'inégalité proposée par Zenga (2007).
- Comme l'ndice de Gini, l'ndice de Zenga prend une valeur entre 0 et 1.
- $L(\alpha)$ est la courbe de lorenz.

Il a été prouvé que l'indice de Zenga est caractérisé par toutes les propriétés principales que toute mesure d'inégalité devrait satisfaire.

2.9 Fonction de bien-être

On considère que la société est formée par une collection de n individus et l'on va s'intéresser à une mesure de bien-être pour l'ensemble de ces n éléments pris comme une entité. La mesure se fera à partir d'une quantité

uni-dimensionnelle que l'on prendra égale soit au revenu, soit à la dépense de consommation que l'on va noter x_i pour l'individu i. On a donc la première donnée de

$$X = (x_1, x_2, ..., x_n) \tag{2.53}$$

qui représente la distribution des revenus (ou de toute autre caractéristique) au niveau de la population. On définit en suite la fonction de bien-être comme une fonction à n arguments :

$$W(x) = V(x_1, ..., x_n). \tag{2.54}$$

Cette fonction a un aspect très normatif et sa construction répond à une série d'axiomes qui précisent les comparaisons que l'on s'autorise à faire entre les individus.

1. **Axiome de Pareto** : la fonction est croissante en chacun de ses termes. On peut affaiblir cette axiome en demandant à la fonction d'être simplement non-décroissante en ses termes. Alors, on peut construire une fonction de bien-être qui restera constante si le revenu des plus riches augmente et ne croîtra que si le revenu des plus pauvres augmente.
2. **Axiome de symétrie ou anonymat** : On doit pouvoir intervertir les individus sans que la valeur de la fonction change. Mais, il existe des problèmes soulevés par la composition des ménages. Les données concernent les ménages, alors que la notion de bien-être s'intéresse aux individus. Il y aura donc une incidence non triviale de la composition des ménages que l'on essayera de gommer au moyen des échelles d'équivalence.
3. **Principe du transfert** : La quasi concavité de la fonction de bien-être implique que si l'on transfère d'un riche vers un pauvre, le bien-être augmente, à condition que le transfert ne change pas l'ordre du classement des individus. Il s'agit du principe dit de Pigou-Dalton.
4. **Autres axiomes** : la littérature économique sur la construction des fonctions de bien-être et des indices d'inégalité est importante. Certains axiomes se recoupent. On peut chercher le nombre minimal d'axiomes qui conduise à la construction de la fonction de bien-être. On consultera à ce propos l'ouvrage de Sen (1997).

La conséquence de ces axiomes, est qu'une fonction de bien-être exprime l'aversion d'une société pour l'inégalité et que cette fonction sera maximale quand tous les ménages auront le même revenu.

2.10 De l'inégalité à la pauvreté

La consultation de la forme de la fonction de bien-être permet de voir que la croissance économique, c'est à dire l'augmentation conjointe de μ et de W peut s'accompagner d'une augmentation des inégalités : certains vont s'enrichir plus vite que les autres. C'est ce que l'on a constaté par exemple au Royaume Uni pendant la période Thatcher. Atkinson (2003) montre comment dans les années 1980, le revenu réel des plus pauvres est resté constant alors que la croissance des revenus a concerné les groupes moyens et surtout les groupes les plus riches. Malgré cela la mesure du bien-être a augmenté.

La pauvreté est ressentie comme un échec et cela justifie que l'on s'y intéresse plus particulièrement. La fonction de bien-être transforme une distribution complète en un nombre permettant d'analyser les effets de mesures de politique économique sur l'ensemble de la distribution des revenus. Si l'on veut concentrer son attention sur les plus pauvres, on va s'intéresser plus particulièrement à une partie de la distribution des revenus, celle qui concerne les plus pauvres, ne serait-ce que pour les compter. On va donc passer de l'analyse des inégalités à l'analyse de la pauvreté en concentrant son attention sur la queue gauche de la distribution des revenus.

Pour concenter son attention sur les plus pauvres, il faut définir ce que l'on appelle une ligne de pauvreté, c'est à dire un seuil en delà duquel une personne (ou un ménage) sera considéré comme pauvre et au delà duquel il basculera dans la catégorie des non-pauvres. On mesure combien ce qu'un tel seuil a d'arbitraire. On peut le définir de deux façons

1. un seuil de pauvreté absolu se définit par rapport à un niveau minimum de subsistance. Le gouvernement Indien par exemple a défini un nombre minimum de calories nécessaires en ville et qui est différent de celui nécessaire à la campagne. En passant par un indice de prix, il arrive à un seuil monétaire de pauvreté en ville et à la campagne. Sur la même base alimentaire, le gouvernement américain a défini un seuil absolu de pauvreté, mais en divisant celui-ci par la part de la nourriture dans le budget d'un ménage moyen. Le RMI(revenu minimum d'insertion) peut également se situer dans ce cadre.
2. Dans les pays développés et plus particulièrement au sein de l'Union Européenne, on préfère définir un seuil relatif de pauvreté. L'Union Européenne a lancé un programme de mesure de la pauvreté où le seuil de pauvreté y est défini par rapport à la moyenne ou la médiane de la distribution des revenus. Sera considéré comme pauvre tout individu touchant un revenu inférieur à 50% ou 60% de la moyenne des revenus de son pays. Il s'agit alors de pauvreté relative, ce qui nous rapproche

de la notion de pauvreté ressentie.

2.11 Les indices de pauvreté

Il existe toute une série d'indices de pauvreté, mais d'une certaine façon les indices les plus simples à comprendre et à manipuler sont les indices linéaires de Foster, Greer, and Thorbecke(1984). Ces indices sont basés sur des moyennes partielles construites à partir de la distribution des revenus. Si $F(.)$ est la distribution des revenus et z le seuil de pauvreté, alors pour un α donné cet indice s'écrit

$$P_\alpha = \int_0^z \left(\frac{z-x}{z}\right)^\alpha dF(x). \tag{2.55}$$

En faisant varier le paramètre α entre 0 et 2, on retrouve un certain nombre des mesures classiques de pauvreté.

- **Pour** $\alpha = 0$, on tombe sur la mesure traditionnelle dite de headcount qui est une mesure de dénombrement :

$$P_0 = \int_0^z f(x)\,dx = F(z). \tag{2.56}$$

  Il s'agit d'une première mesure, intéressante en soi, elle permet de connaître le nombre de pauvres en multipliant simplement P_0 par la taille de la population. Toutefois cette mesure est insuffisante car elle ne distingue pas entre les pauvres et ne tient pas compte de leur niveau de pauvreté, c'est à dire qu'elle ne distingue pas entre les individus qui sont près de la ligne de pauvreté et ceux qui en sont loin.
- **Pour** $\alpha = 1$, on tombe sur une mesure faisant intervenir le déficit de pauvreté ou poverty gap $z - x_i$ qui affecte chaque individu en dessous du seuil de pauvreté :

$$P_1 = \int_0^z (1 - x/z)\, f(x)\,dx$$

  Cet indice respecte le principe de transfert à l'inverse de la mesure de comptage P_0 qui ne le respecte pas. Cet indice est continu alors que le head count ne l'est pas. Mais il est insensible à certains types de transferts entre les pauvres.
- **Pour** $\alpha = 2$, on arrive à une mesure qui est sensible à la distribution parmi les pauvres :

$$P_2 = \int_0^z (1 - x/z)^2 f(x)\,dx, \tag{2.57}$$

mais qui n'est pas très souvent employée. Atkinson (1987) examine simplement les propriétés d'une généralisation des deux premiers indices et leur relation avec la dominance stochastique restreinte, notion que l'on explicitera plus bas.

L'indice de Foster, Greer, and Thorbecke (1984) répond à la propriété de décomposabilité car il a une structure linéaire. Considérons par exemple une partition de la population entre urbaine et rurale. Si X représente l'ensemble des revenus, la partition de X se définira comme $X = X_U + X_R$. Appelons p la proportion de X_U dans X. Alors l'indice total de pauvreté se décomposera en

$$\begin{aligned} P_\alpha &= p\int_0^z \left(\frac{z-x}{z}\right)^\alpha dF(x_U) + (1-p)\int_0^z \left(\frac{z-x}{z}\right)^\alpha dF(x_R) \quad (2.58) \\ &= pP_\alpha^U + (1-p)P_\alpha^R. \end{aligned}$$

Une autre classe d'indices de pauvreté a été proposée à la suite de Sen (1976) pour tenir compte de l'inégalité entre les pauvres. L'indice original de Sen, P_s, combine une mesure de headcount P_0, une mesure du poverty gap,

$$zP_0I_p = \int_0^z (z-x)\, f(x)\, dx$$

et un indice de Gini G_p calculé sur le segment $x < z$. Cet indice se note

$$P_s = P_0\left(I_P + (1-I_P)\, G_p\right) = P_0\left(1 - (1-I_P)\,\frac{\mu_p}{z}\right) \quad (2.59)$$

où μ_p est la moyenne des revenus parmi les pauvres :

$$\mu_p = \int_0^z xf(x)\, dx/F(z) \quad et \quad I_p = 1 - \frac{\mu_p}{z}. \quad (2.60)$$

Quand il n'y a pas d'inégalité entre les pauvres ($G_p = 0$), alors $P_s = P_1$. Quand l'inégalité devient extrême ($G_p = 1$), on retombe sur la mesure de headcount, ce que traduit bien la factorisation

$$P_s = P_0G_p + P_1(1-G_p). \quad (2.61)$$

Mais tout comme l'indice de Gini, cet indice n'est pas décomposable. Il viole également l'axiome de transfert et de plus n'est pas continu. Shorrocks (1995) a proposé une modification de cet indice appelé aussi l'indice de Sen-Schorrocks-Thon qui résoud une partie de ces difficultés. Cet indice s'écrit par analogie avec l'indice de Sen comme

$$P_{SST} = (2-P_0)\, P_0I_p + P_0^2\,(1-I_p)\, G_p. \quad (2.62)$$

2.12 Pauvreté et inégalité

La formulation initiale de la fonction de bien-être (2.54) implique qu'une augmentation du bien-être peut tout à fait s'accompagner d'un accroissement des inégalités. Comment inclure dans cette décomposition une attention plus particulière à la pauvreté ? En d'autres termes, quelle forme doit-on considérer pour $W(x)$ si l'on veut maximiser le bien-être tout en insistant sur la pauvreté. Atkinson (1987) traite de cette question dans la section 3 de son papier en distinguant quatre options possibles.

- La première option consiste à ne pas se soucier particulièrement de la pauvreté. On va simplement maximiser

$$W(x) = \mu(1 - I), \tag{2.63}$$

où I est un indice d'inégalité et μI mesure le coût de l'inégalité. Si l'on a choisi de manière adéquate la fonction de bien-être, on peut décomposer cet indice en distinguant le groupe des pauvres du reste de la population. On pourra donc mesurer l'évolution de la pauvreté sans avoir la réduction de la pauvreté comme objectif principal.

- Dans une deuxième option, on va chercher à introduire un coût prioritaire pour la pauvreté $C_p = \mu P$ et laissant un rôle secondaire au coût de l'inégalité. Ceci peut se faire en adoptant une fonction de bien-être du type

$$W(x) = \mu - \mu P - \mu I.$$

Atkinson (1987) indique que dans ce cas, il est logique d'utiliser une mesure de comptage pour P et une mesure satisfaisant le principe de transfert pour I.

- La troisième option consiste à se focaliser uniquement sur la pauvreté. La fonction de bien-être à maximiser sera alors de la forme

$$W(x) = \mu - \mu P. \tag{2.64}$$

- Enfin, la dernière option consiste à utiliser un arbitrage entre inégalité et pauvreté. On aura toujours la fonction de bien-être donnée en

$$W(x) = \mu - \mu I - \mu P. \tag{2.65}$$

Mais cette fois-ci des considérations de justice vont conduire à utiliser pour I un indice de Gini calculé sur toute la population et pour P un indice de pauvreté de Sen (1976) modifié.

Ces considérations montrent que la construction de la fonction de bien-être peut être relativement complexe quant à ses propriétés et à la façon dont sont agrégés les individus. La forme simple était peut-être un peu trop simple.

2.13 La décomposition des indices

Certains indices d'inégalité et de pauvreté peuvent facilement se décomposer comme on la vu avec les indices de Foster, Greer, et Thorbecke (1984). La décomposition se fait alors par groupes de la population. Mais dans cette procédure aucune explication n'est donnée en fonction des caractéristiques des sous groupes. Oaxaca (1973) a le premier tenté une explication des inégalités en utilisant une technique de régression.

Oaxaca (1973) s'intéresse aux inégalités de salaire entre hommes et femmes. Pour chaque groupe, on estime une équation de salaire

$$\log(W_i) = X_i\beta + u_i, \quad i = h, f. \tag{2.66}$$

On regarde ensuite les différences salariales moyennes entre hommes et femmes. Une partie de cette différence s'explique par des différences de caractéristiques objectives mesurées par X_h, X_f, l'autre partie s'explique par des différences de rendements de ces mêmes caractéristiques, c'est à dire par la discrimination que le marché opère entre hommes et femmes. Comme dans b une régression $log(Wi) = \overline{X}_i\widehat{\beta}_i$, on aura la décomposition suivante appelée déomposition de Oaxaca :

$$\log(W_h) - \log(W_f) = (X_h - X_f)\,\widehat{\beta}_h + \overline{X}_f\left(\widehat{\beta}_h - \overline{\beta}_f\right). \tag{2.67}$$

Ce type de décomposition a donné lieu à des développements importants dans la littérature. Par exemple Juhn, Murphy, and Pierce (1993) généralisent le résultat précédent aux différents quantiles de la distribution des résidus. Radchenko et Yun (2003) donnent une approche Bayésienne qui permet d'implémenter facilement des tests de significativité.

La décomposition de Oaxaca (1973) est b asée sur une hypothèse de régression linéaire. Yun (2004) en donne une généralisation non-linéaire qui permet de proposer une décomposition des mesures de head count. Celles-ci étant assimilables à des proportions, on peut les relier à des variables explicatives au moyen d'un modèle probit. Une des applications dans Yun (2004) concerne justement les modèles probit. Une équation de régression par groupe permet d'expliquer le ratio entre la dépense y et le seuil de pauvreté z :

$$\log\left(\frac{y}{z}\right) = X\beta + e, \tag{2.68}$$

où X est une matrice de caractéristiques personnelles à k composantes pour n observations. Sous une hypothèse de normalité pour e, la probabilité d'être pauvre pour le groupe des n individus est égale au vecteur $\Phi(-X\beta/\sigma)$ où σ^2 est la variance des résidus. Quand n tend vers l'infini, la moyenne des

$\Phi(-X\beta/\sigma)$tend vers le head count ratio, c'est à dire P_0 dans nos notations. Alors, on aura la décomposition suivante de la différence entre deux indices de pauvreté correspondant à deux groupes distincts A et B :

$$P_A^0 - P_B^0 = \left[\overline{\Phi(-X_A\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_A/\sigma_A)}\right] + \left[\overline{\Phi(-X_B\beta_A/\sigma_A)} - \overline{\Phi(X_B\beta_B/\sigma_B)}\right] \tag{2.69}$$

ce qui correspond à la différence entre les caractéristiques et la différence entre les coefficients.

L'application de la procédure de Yun (2004) permet de pondérer cette décomposition en fonction du poids de chacune des k caractéristiques individuelles

$$\begin{aligned} P_A^0 - P_B^0 &= \sum_{i=1}^{k} W_{\Delta x}^i \left[\overline{\Phi(-X_A\beta_A/\sigma_A)} - \overline{\Phi(-X_B\beta_A/\sigma_A)}\right] \\ &+ \sum_{i=1}^{k} W_{\Delta\beta}^i \left[\overline{\Phi(-X_B\beta_A/\sigma_A)} - \overline{\Phi(X_B\beta_B/\sigma_B)}\right] \end{aligned} \tag{2.70}$$

où les poids $\sum_{i=1}^{k} W_{\Delta x}^i$ et $\sum_{i=1}^{k} W_{\Delta\beta}^i$ sont donnés dans Bhaumik, Gang, et Yun (2006) sur un argument de linéarisation de Yun (2004).

Chapitre 3: Estimation et Simulation

Chapitre 3

Estimation et Simulation

3.1 Introduction

Mesurer et analyser les revenus, les pertes, les risques et d'autres résultats aléatoires, que nous dénotions par X, a été un domaine de recherche actif et fructueux, notamment dans les économétrie et science actuarielle. L'indice de *Gini* est sans doute la mesure la plus populaire d'inégalité, avec un certain nombres d'extensions et de généralisations disponibles dans la littérature. Gardant à l'esprit que les notions de pauvres et riches sont les uns par rapport aux autres, *Zenga* construit un indice qui reflète cette relativité.

3.2 Estimation de l'indice de Zenga

3.2.1 Rappel sur l'indice de Zenga

Soit $F(x) = P[X \leq x]$ la fonction de distribution cumulative (fdc) de la variable aléatoire X, et

$$Q(t) = \inf\{x : F(x) \geq t\}, \quad t \in [0\ ;\ 1]$$

dénote la fonction quantile correspondante.

L'indice Z_F de l'inégalité de Zenga est défini par la formule

$$Z_F = \int_0^1 z_F(p)dp \tag{3.1}$$

où $z_F(p)$ est la courbe de Zenga, donnée par :

$$z_F(t) = 1 - \frac{L_F(p)}{p}.\frac{1-p}{1-L_F(p)} \tag{3.2}$$

$L(p)$ est la fonction de Lorenz donnée par

$$L(p) = \frac{1}{\mu} \int_0^P Q(s)ds. \tag{3.3}$$

3.2.2 Estimateur traditionnel de l'indice Zenga

Soit $X_1, X_2, ..., X_n$ observations indépendants identiquement distribuées suivant de la fonction de distribution F. On considère l'estimation empirique de F par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}} \tag{3.4}$$

avec 1_A désigne la fonction d'indicateur sur l'ensembe A. Par le remplacement de l'estimateur (3.4) dans la formule (3.1), nous arrivons à un estimateur empirique de l'indice de Zenga traditionnel (par exemple, Greselin et Pasquazzi 2009 ; Greselin et al. 2013)

$$\hat{Z}_n = 1 - \frac{1}{n} \sum_{i=1}^{n-1} \frac{i^{-1} \sum_{k=1}^{i} X_{k:n}}{(n-i)^{-1} \sum_{k=i+1}^{n} X_{k:n}} \tag{3.5}$$

où $X_{1:n} < X_{2:n} < ... < X_{n:n}$ sont les statistiques d'ordre basées sur la série $X_1, X_2, ..., X_n$.

Théorème 3.1 *Si X est une v.a. de CDF F avec le moment $E[X^{2+\varepsilon}]$ est fini pour tout $\varepsilon > 0$. Alors on a la représentation asymptotique*

$$\sqrt{n}(\hat{Z}_n - Z_F) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) + o_p(1) \tag{3.6}$$

ici $o_p(1)$ désigne une variable aléatoire qui converge vers 0 en probabilité lorsque $n \to \infty$, et

$$h(X_i) = \int_0^{\infty} \left(1_{\{X_i \leq x\}} - F(x)\right) \omega_F(F(x))dx$$

avec la fonction de poids

$$\omega_F(t) = \frac{1}{\mu_F} \int_0^t \left(\frac{1}{p} - 1\right) \frac{L_F(p)}{(1 - L_F(p))^2} dp + \frac{1}{\mu_F} \int_t^0 \left(\frac{1}{p} - 1\right) \frac{L_F(p)}{1 - L_F(p)}. \tag{3.7}$$

Preuve: [Voir Greselin et al. 2013] ■

3.3 Indice de Gini

En général, l'indice de Gini est une fonction $G : \mathbb{R}_n^+ \longrightarrow [0; 1]$ qui attribue à chaque vecteur de revenu non négatif un nombre réel entre 0 et 1, ce qui représente le niveau d'inégalité de la société. Cette mesure est 0 en égalité maximale et 1 en parfaite inégalité. Par conséquent, l'indice de Gini peut être exprimé comme

$$G = 2 \int_0^1 (p - L(p))dp, \tag{3.8}$$

tel que $p = F(x)$ est une fonction de distribution cumulative (fdc) de non-négatif revenu avec espérance positive.

En utilisant la définition de l'indice de Gini dans l'équation (3.8), G est la distance entre la ligne d'égalité et la courbe de Lorenz, et en appliquant un changement de variable $p = F(x)$, il peut être trouvé que :

$$G = \frac{2}{\mu} \int_0^{+\infty} xF(x)dF(x) - 1. \tag{3.9}$$

avec μ représente l'espérance de la v.a. X. Supposons qu'un échantillon i.i.d de taille n soit tiré aléatoirement de la population, et $\hat{F}_n$ désigne la fonction de distribution empirique correspondante. Soit $X_1, ..., X_n$ un échantillon aléatoire et $X_{1:n} \leq ... \leq X_{n:n}$ les statistiques d'ordre obtenues à partir de l'échantillon $X_1, ..., X_n$. Ensuite, un estimateur empirique de μ est donné par

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{3.10}$$

Un estimateur alternatif de l'indice de Gini peut être obtenu par le remplacement de la cdf empirique $\hat{F}_n$ de revenu au lieu de sa distribution correspondante fonction F et de remplacer $\hat{\mu}_n$ à la place de μ dans (3.9), comme suit :

$$\hat{G}_n = \frac{2}{\hat{\mu}_n} \int_0^{+\infty} x\hat{F}_n(x)d\hat{F}_n(x) - 1. \tag{3.11}$$

À cet égard, l'exemple de l'indice de Gini peut être exprimé comme :

$$\hat{G} = \frac{1}{\hat{\mu}_n} \int_0^{+\infty} xd(\hat{F}(x))^2 - 1 = \frac{2\sum_{i=1}^{n} X_{i:n}(i - \frac{1}{2})}{n\sum_{i=1}^{n} X_i} - 1. \tag{3.12}$$

3.4 Estimation semi paramétrique

3.4.1 Indice de Zenga

Soit F la fonction de distribution cumulative (fdc) de la population représentée par une variable aléatoire non négative $X \geq 0$. Nous supposons que F est continue et strictement croissante. L'inverse (généralisé) $Q : (0,1) \to [0,\infty)$ du fdc F, connu dans la littérature comme la fonction quantile, est défini pour tout $t \in (0,1)$ par la formule

$$Q(t) = \inf\{x : F(x) \geq t\}$$

Les attentes de queue conditionnelle supérieure et inférieure sont $E[X|X > Q(t)]$ et $E[X|X \leq Q(t)]$, respectivement. Puisque F est continu, ils coïncident avec le haut et le bas valeurs de queue à risque (par exemple, Denuit et al., 2005) :

$$Z = 1 - \int_0^1 \left(\frac{1-t}{t} \frac{\int_0^t Q(s)ds}{\int_t^1 Q(s)ds} \right) dt \tag{3.13}$$

et

$$\tilde{Z}_n = 1 - \int_0^1 \left(\frac{1-t}{t} \frac{\int_0^t Q_n(s)ds}{\int_t^1 Q_n(s)ds} \right) \tag{3.14}$$

3.4.2 Estimateur à queue lourde de l'indice de Zenga

Puisque nous sommes concernés par les populations à queue lourde, nous travaillons inévitablement avec les fdc qui varient régulièrement à l'infini. Par conséquent, nous supposons que F satisfait

$$\lim_{t \longrightarrow \infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma} \tag{3.15}$$

pour certains $\gamma > 0$, ce que l'on appelle l'indice de la queue de ladistribution.

Notre estimateur à queue lourde de l'indice Z_F est basé sur l'estimateur de Weissman (1978) de l'estimateur de Hill-hautiles et de Hill's (1975) de l'indice de queue γ. A savoir, avec une suite d'entiers $k := k_n \to \infty$ tels que $k/n \to 0$ quand $n \to \infty$, le Hill's l'estimateur :

$$\hat{\gamma}_n = \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right). \tag{3.16}$$

De Haan (1994) a étudié les propriétés asymptotiques de l'estimateur. Avec la fonction quantile empirique définie par

$$Q_n(t) = \inf\{x : F_n(x) \geq t\},$$

Z est estimé par :

$$\bar{Z}_{n,k} = 1 - \int_0^1 \left(\frac{1-t}{t} \frac{\int_0^t Q_n(s)ds}{\int_t^{1-k/n} Q_n(s)ds + \frac{kX_{n-k:n}}{n(1-\hat{\gamma}_n)}} \right) dt. \tag{3.17}$$

Théorie de l'inférence

Établir les distributions asymptotiques des estimateurs dans les situations à queue lourde nécessite une hypothèse plus forte que celle de l'équation (3.15), ce qui est suffisant pour la cohérence de résultats. Par conséquent, à partir de maintenant, nous supposons que F satisfait le condition généralisé de variation régulière de second ordre avec le paramètre de second ordre $\rho \leq 0$, qui signifie qu'il existe une fonction α sur $[0, \infty[$ telle que

- $\alpha(t) \to 0$ lorsque $t \to \infty$;
- $\alpha(t)$ ne change pas de signe pour tout t suffisamment grand ; et
- l'équation

$$\lim_{t \to +\infty} \frac{1}{\alpha(t)} \left(\frac{1-F(tx)}{1-F(t)} \right) - x^{-1/\gamma} = x^{-1/\gamma} \frac{x^{\rho/\gamma} - 1}{\rho/\gamma} \tag{3.18}$$

est vérifiée pour tout $x > 0$, avec le côté droit interprété comme $x^{-1/\gamma} \log x$ quand $\rho = 0$.

Théorème 3.2 *Supposons que le fdc F vérifie la condition (3.18) avec un $\gamma \in (1/2, 1)$ et $\rho \leq 0$, et soit $k = k_n \to \infty$ quand $n \to \infty$ est tel que $k/n \to 0$ et $\sqrt{k}\alpha(Q(1-k/n)) \to 0$. Puis sur un espace de probabilité approprié, et avec des ponts Brownien construits de manière appropriée $\mathcal{B}_n$, nous avons*

$$\begin{aligned} \frac{\sqrt{n}(\bar{Z}_{n,k} - Z)}{\sqrt{k/n}Q(1-k/n)} &= -\int_0^{1-k/n} \frac{\mathcal{B}_n(s)v(s)}{\sqrt{k/n}Q(1-k/n)} dQ(s) \\ &\quad + \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2} \sqrt{\frac{n}{k}} \mathcal{B}_n\left(1-\frac{k}{n}\right) \\ &\quad - \frac{\gamma v(1-k/n)}{(1-\gamma)^2} \sqrt{\frac{n}{k}} \int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s} ds + op(1) \end{aligned} \tag{3.19}$$

lorsque $n \longrightarrow \infty$, où, pour $0 \leq s \leq 1$,

$$v(s) = \int_0^s \frac{1-t}{t} \frac{\int_0^t Q(s)ds}{\int_t^1 Q(s)ds} dt. \tag{3.20}$$

Corollaire 3.1 *Sous les hypothèses du théorème 3.2 , nous avons*

$$\frac{\sqrt{n}(\bar{Z}_{n,k} - Z)}{\sqrt{k/n}Q(1-k/n)} \longrightarrow \mathcal{N}(0, \sigma_Z^2) \quad . \tag{3.21}$$

avec

$$\sigma_Z^2 = \frac{\gamma^4}{(1-\gamma)^4(2\gamma-1)} v^2(1). \tag{3.22}$$

3.4.3 Indice de Gini

Les Mêmes étapes utilisées dans Zenga, sont utilisées dans Gini, défini par

$$G = 1 - \frac{2}{\mu} \int_0^1 \int_0^t Q(s)ds. \tag{3.23}$$

avec la fonction quantile empirique nous estimons G par

$$\bar{G}_{n,k} = 1 - \frac{2}{\hat{\mu}_n} \int_0^1 \int_0^t Q_n(s)dsdt \tag{3.24}$$

avec

$$\begin{aligned} \hat{\mu}_n &= \int_0^{1-k/n} Q_n(s)ds + \int_{1-k/n}^1 Q_n^w(s)ds \\ &= \int_0^{1-k/n} Q_n(s)ds + \frac{\frac{k}{n}X_{n-k:n}}{1-\hat{\gamma}_n} \end{aligned} \tag{3.25}$$

Théorème 3.3 *Supposons que le fdc F vérifie la condition (3.18) avec un $\gamma \in (1/2, 1)$ et $\rho \leq 0$, et soit $k = k_n \longrightarrow \infty$ quand $n \longrightarrow \infty$ est tel que $k/n \longrightarrow 0$ et $\sqrt{k}\alpha(\mathbb{Q}(1-k/n)) \longrightarrow 0$. Puis sur un espace de probabilité approprié, et avec ponts Brownien $\mathcal{B}_n$,construits de manière appropriée nous avons*

$$\begin{aligned} \frac{\sqrt{n}(\bar{G}_{n,k} - G)}{v\sqrt{k/n}\mathbb{Q}(1-k/n)} &= -\int_0^{1-k/n} \frac{v\mathcal{B}_n(s)}{\sqrt{k/n}Q(1-k/n)} dQ(s) \\ &+ \frac{\gamma^2 v}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right) \\ &- \frac{\gamma v}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s} ds + op(1) \end{aligned} \tag{3.26}$$

lorsque $n \to \infty$, *où,*

$$\upsilon = \frac{2}{\mu^2} \int_0^1 \int_0^t Q(s)ds.$$

Corollaire 3.2 *Sous lesmêmes hypothèses du théorème (3.3). On a*

$$\frac{\sqrt{n}(\bar{G}_{n,k} - G)}{\sigma(\gamma)\sqrt{k/n}\mathbb{Q}(1-k/n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \text{ comme } n \longrightarrow \infty$$

où

$$\sigma^2(\gamma) = \frac{\upsilon^2\gamma^4}{(1-\gamma)^4(2\gamma-1)}.$$

Preuve: Notons $U_i = F(X_i)$ pour$i = 1, 2, ..., n$. Alors $U_1, U_2, ..., U_n$ est une séquence de i.i.d. variables aléatoires suivant la distribution uniforme sur $[0, 1]$. Ce qui suit théorème montre que est les processus empiriques et quantiles basés sur la séquence $U_1, U_2, ..., U_n$ peut être approchés par une série de ponts browniens ; (voir Csörgő et Horváth 1993)

$$\begin{aligned}
\bar{G}_{n,k} - G &= \left(1 - \frac{2}{\hat{\mu}_n}\int_0^1\int_0^t Q_n(s)dsdt\right) - \left(1 - \frac{2}{\mu}\int_0^1\int_0^t Q(s)dsdt\right) \quad (3.27)\\
&= -\frac{2}{\hat{\mu}_n}\int_0^1\int_0^t Q_n(s)dsdt + \frac{2}{\mu}\int_0^1\int_0^t Q(s)dsdt\\
&= -\frac{2}{\hat{\mu}_n}\int_0^1\int_0^t Q_n(s)dsdt + \frac{2}{\hat{\mu}_n}\int_0^1\int_0^t Q(s)dsdt\\
&\quad + \frac{2}{\mu}\int_0^1\int_0^t Q(s)dsdt - \frac{2}{\hat{\mu}_n}\int_0^1\int_0^t Q(s)dsdt
\end{aligned}$$

$$\bar{G}_{n,k} - G = A_n + B_n \quad (3.28)$$

où

$$A_n = -\frac{2}{\hat{\mu}_n}\left(\int_0^1\int_0^t [Q_n(s) - Q(s)]\, dsdt\right) \quad (3.29)$$

et

$$\begin{aligned}
B_n &= \left(\frac{2}{\mu} - \frac{2}{\hat{\mu}_n}\right)\int_0^1\int_0^t Q(s)dsdt \quad (3.30)\\
&= 2\left(\frac{(\hat{\mu}_n - \mu)}{\mu\hat{\mu}_n}\right)\int_0^1\int_0^t Q(s)dsdt
\end{aligned}$$

A_n qui fait partie intégrante du processus quantile général $Q_n - Q$. Pour le réduire à une intégrale de l'empirique générale processus $F_n - F$, nous utilisons le procédé Vervaat (général)

$$V_n(t) = \int_0^t (Q_n(s) - Q(s))ds + \int_{-\infty}^{Q(t)} (F_n(x) - F(x))dx \tag{3.31}$$

Le processus $V_n(t)$ vérifie les conditions aux limites $V_n(0) = 0$ et $V_n(1) = 0$, est non négatif pour tout $t \in [0, 1]$, et tel que

$$\sqrt{n}V_n(t) \leq |e_n(t)||Q_n(t) - Q(t)|. \tag{3.32}$$

Par conséquent, en rappelant que $e_n(t) = \sqrt{n}(F_n(Q(t)) - t)$ nous concluons à partir de l'équation(3.31) que la différence entre les quantités

$$\sqrt{n}\int_0^t (Q_n(s) - Q(s))ds \tag{3.33}$$

et

$$-\sqrt{n}\int_{-\infty}^{Q(t)} (F_n(x) - F(x))dx \tag{3.34}$$

tend vers zéro quand $n \to \infty$ quand $Q_n(t)$ converge vers $Q(t)$, ce qui est vrai F est continu et strictement croissant. C'est l'idée même d'employer le Vervaat processus dans la présente preuve, car il nous permet de remplacer la quantité (3.33) par (3.34) ce qui est beaucoup plus facile à aborder. Nous avons l'équation suivante

$$A_n(t) = \int_0^{1-k/n} (Q_n(s) - Q(s))ds - \int_0^t (Q_n(s) - Q(s))ds \tag{3.35}$$

$$= -\int_{Q(t)}^{Q(1-k/n)} (F_n(x) - F(x))dx + V_n(1 - k/n) - V_n(t). \tag{3.36}$$

que nous appliquons sur Eq.(3.28). En changeant la variable de l'intégration, nous obtenons

$$A_n(t) = -\int_t^{1-k/n} \frac{e_n(s)}{\sqrt{n}} dQ(s) + V_n(1 - k/n) - V_n(t)$$

et

$$A_n(0) - A_n(t) = -\int_0^t \frac{e_n(s)}{\sqrt{n}} dQ(s) + V_n(t). \tag{3.37}$$

Alors

$$\frac{\sqrt{n}A_n(t)}{\sqrt{k/n}Q(1 - k/n)} = -\frac{\int_{Q(t)}^{Q(1-k/n)} e_n(F(x))dx}{\sqrt{k/n}Q(1 - k/n)} + O_p\frac{| e_n(1 - k/n) ||}{\sqrt{k/n}Q(1 - k/n)}.$$

D'après le résultat de Peng L. 2001, Necir, Rassoul and Zitikis (2010), il existe une suite des ponts Brownian $\{B_n(s), 0 \leq s \leq 1\}_{n\geq 1}$ telle que, pour tout n assez grand, on a :

$$\begin{aligned}\frac{\sqrt{n}(\hat{\mu}_n-\mu)}{\sqrt{k/n}Q(1-k/n)} \overset{d}{=} & -\int_0^{1-k/n} \frac{B_n(s)}{\sqrt{k/n}Q(1-k/n)}dQ(s) \\ & + \left\{\sqrt{n/k}\mathcal{B}_n(s)B_n(1-k/n)\right\} \\ & -\sqrt{n/k}\int_{1-k/n}^1 \frac{B_n(s)}{1-s}ds + o_P(1),\end{aligned}$$

Alors

$$\frac{\sqrt{n}(\bar{G}_{n,k}-G)}{\sqrt{k/n}Q(1-k/n)} = \sum_{i=1}^{3} T_{n,i} + o_p(1)$$

où

$$T_{n,1} = -\int_0^{1-k/n} \frac{\mathcal{B}_n(s)v}{\sqrt{k/n}X_{n-k:n}}dQ(s)$$

$$T_{n,2} = \frac{\gamma^2 v}{(1-\gamma)^2}\sqrt{\frac{k}{n}}\mathcal{B}_n(s)(1-\frac{k}{n})$$

$$T_{n,3} = -\frac{\gamma v}{(1-\gamma)^2}\sqrt{\frac{k}{n}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s}ds$$

la somme $T_{n,1}+T_{n,2}+T_{n,3}$ est une variable aléatoire Gaussienne centrée. Pour calculer sa variance asymptotique nous établissons la limite suivante

$$\begin{aligned}E[T_{n,1}^2] &\to \frac{2\gamma}{2\gamma-1}, \quad E[T_{n,2}^2] \to \frac{\gamma^4}{(1-\gamma)^4} \\ E[T_{n,3}^2] &\to \frac{\gamma^2}{(1-\gamma)^4}, \quad E[T_{n,1}T_{n,2}] \to \frac{\gamma^2}{(1-\gamma)^2} \\ E[T_{n,1}T_{n,3}] &\to \frac{\gamma}{(1-\gamma)^2}, E[T_{n,2}T_{n,3}] \to \frac{\gamma^3}{(1-\gamma)^4}.\end{aligned}$$

■

3.5 Simulations

Partie 1 : Pour montrer le comportement de l'estimateur empirique de l'indice de Gini, on réalise quelques simulations des échantillons de taille n deux distribution, Pareto et Fréchet pour trois valeurs de l'indice $\gamma \in \{0.25, 0.5, 0.75\}$, on choisit la taille selon les valeurs n = 100, 500,1000 et 2000. Ensuite, nous avons calculé la valeurs de l'estimateur de l'indice de

Gini pour chaque loi, et on compare avec les valeurs théoriques, et on calcule l'erreur. Les différents résultats de simulation sont présentés dans les tableaux (3.1,3.2,3.3,3.4,3.5 et 3.6) suivants :

avec $erreur = |GP - G_m|$

GP est théorie (l'integrale) et G_m est empirique (la somme)

TAB. 3.1 – Simulation de l'indice de Gini par la loi de Pareto pour $\gamma = 0.25$

n	100	500	1000	2000
GP	0.1428572	0.1428572	0.1428572	0.1428572
G_{em}	0.1748046	0.1463	0.1448	0.14334
erreur	0.01394739	0.0036	0.00194	0.0004

TAB. 3.2 – Simulation de l'indice de Gini par la loi de Pareto pour $\gamma = 0.5$

n	100	500	1000	2000
GP	0.3268	0.3268	0.3268	0.3268
G_{em}	0.35207	0.33383	0.3298	0.327293
erreur	0.0252	0.006	0.002942	0.000425

TAB. 3.3 – Simulation de l'indice de Gini par la loi de Pareto pour $\gamma = 0.75$

n	100	500	1000	2000
G	0.555803	0.555803	0.555803	0.555803
G_{em}	0.623548	0.53429	0.5072	0.49509
erreur	0.6774	-0.021513	-0.04854	-0.0607

Partie 02 : Pour montrer le comportement de l'estimateur semiparamétrique de l'indice de Gini, on réalise quelques simulations des échantillons de taille n deux distribution, Pareto et Fréchet pour deux valeurs pour l'indice $\gamma \in \{2/3, 3/4\}$, on choisit la taille selon les valeurs n = 100, 500,1000 et 2000. Ensuite, nous avons estimé la fraction optimale k_{opt} et l'estimateur de Hill $\hat{\gamma}_n$, et la valeurs de l'estimateur de l'indice de Gini pour chaque loi, et on compare avec les valeurs théoriques, et on calcule l'erreur. Les différents résultats de simulation sont présentés dans les tableaux (3.7,3.8,3.9 et 3.10) suivants :

TAB. 3.4 – Simulation de l'indice de Gini par la loi de Fréchet pour $\alpha = 0.25$

n	100	500	1000	2000
GF	0.1892072	0.1892072	0.1892072	0.1892072
G_{em}	0.19433	0.199732	0.195737	0.1909272
erreur	0.020006	0.01052	0.006532	0.0017196

TAB. 3.5 – Simulation de l'indice de Gini par la loi de Fréchet pour $\alpha = 0.5$

n	100	500	1000	2000
GF	0.4142136	0.4142136	0.4142136	0.4142136
G_{em}	0.4462247	0.4358977	0.4315344	0.4307771
erreur	0.0320113	0.021684	0.01732	0.016563

TAB. 3.6 – Simulation de l'indice de Gini par la loi de Fréchet pour $\alpha = 0.75$

n	100	500	1000	2000
GF	0.68179	0.68179	0.68179	0.68179
G_{em}	0.7421	0.7207	0.7062	0.68417
erreur	0.0611	0.03896	0.024489	0.00238

TAB. 3.7 – Simulation de l'estimateur semi paramétrique de l'indice de Gini par la loi de Pareto pour $\gamma = 3/4$

n	500	1000	2000
Kop	25	55	29
$\hat{\gamma}_n$	0.765	0.66364	0.6155
$\hat{G}$	0.5219	0.5276	0.53342
erreur	0.0339	0.0282	0.0228

TAB. 3.8 – Simulation de l'estimateur semi paramétrique de l'indice de Gini par la loi de Pareto pour $\gamma = 2/3$

n	500	1000	2000
Kop	26	45	95
$\hat{\gamma}_n$	0.6481	0.6407417	0.66591
$\hat{G}$	0.42872	0.479377	0.47442
erreur	0.047 19	0.00347	0.00149

TAB. 3.9 – Simulation de l'estimateur semi paramétrique de l'indice de Gini par la loi de Fréchet pour $\gamma = 3/4$

n	500	1000	2000
kop	26	50	93
$\hat{\gamma}_n$	0.7677	0.73811	0.88098
$\hat{G}$	0.63408	0.6409377	0.69456
erreur	0.047899	0.0408523	0.01277

TAB. 3.10 – Simulation de l'estimateur semi paramétrique de l'indice de Gini par la loi de Fréchet pour $\gamma = 2/3$

n	500	1000	2000
kop	27	51	91
$\hat{\gamma}_n$	0.7209	0.7977	0.68644
$\hat{G}$	0.48598	0.5061	0.5187
erreur	0.07882	0.0584	0.0458

Introduction générale

Conclusion générale

Dans ce travail, nous avons étudié et approxime des indicateurs statistiques dans l'économie qui ont un impact significatif dans le monde d'aujourd'hui. Mesurer les changements dans la répartition des revenus et l'inference économique est un objectif majeur de la recherche sur l'inégalité. Le développement de l'inégalité économique, une préoccupation majeure pour une croissance énorme avec des niveaux élevés de revenu, la raison est de considérer les mesures traditionnelles de l'inégalité et de fournir de nouvelles façons de collecter la distribution des fluctuations du revenu complet. L'objectif principal de notre approche utiliser est d'obtenir des résultats pour certaine inférences statistiques des quelques indices en économie (indice de Gini et indice de Zenga).

Nous avons identifié quelques variables aléatoires et des indicateurs généraux, et nous avons montré que les indicateurs économiques sont une classe de mesures connexes, et nous avons mené une étude les estimations pour indice Gini et indice de Zenga et proposer un nouvelle estimateur de l'indice Gini, les procédures détaillées et les exemples numériques montrent également comment utiliser l'indice de Gini par les lois de Pareto et Fréchet. Cette proposition est peut-être concluante ils sont truqués lors de distributions supplémentaires pour des considérations théoriques dans le compte.

Bibliographie

[1] Escoufier, Y. (1973). Le traitement des variables vectorielles. Biometrics, 751-760.

[2] Plya, G. (1930). Sur quelques points de la théorie des probabilités. Ann. Inst. H. Poincaré, 1(2), 117-161.

[3] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

[4] Trouvelot, A., Kough, J. L., & Gianinazzi-Pearson, V. (1986). Mesure du taux de mycorhization VA d'un système radiculaire. Recherche de méthodes d'estimation ayant une signification fonctionnelle. In Physiological and genetical aspects of mycorrhizae= Aspects physiologiques et genetiques des mycorhizes : proceedings of the 1st European Symposium on Mycorrhizae, Dijon, 1-5 July 1985. Paris : Institut national de le recherche agronomique, c1986..

[5] Mussard, S. (2006). La décomposition des mesures d'inégalité en sources de revenu : l'indice de Gini et les généralisations. Cahier de recherche/Working Paper, 6, 05.

[6] Gini C., 1912, Variabilità e mutabilità, Bologna, Italy.

[7] Fougère, D., & Kramarz, F. (2001). La mobilité salariale en France de 1967 à 1999. Inégalités économiques, 333-354.

[8] Hourriez, J. M., & Roux, V. (2001). Une vue d'ensemble des inégalités de revenu et de patrimoine. Inégalités économiques, Rapport pour le Conseil d'Analyse économique, La Documentation Française, Paris, 269.

[9] Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. Econometrica : journal of the econometric society, 761-766.

[10] Marciniak, S. J., Yun, C. Y., Oyadomari, S., Novoa, I., Zhang, Y., Jungreis, R., ... & Ron, D. (2004). CHOP induces death by promoting protein synthesis and oxidation in the stressed endoplasmic reticulum. Genes & development, 18(24), 3066-3077.