http://alain.pichereau.pages.perso-orange.fr
marc.pichereau@wanadoo.fr

Concentration d'une série.

Résumé

Le point de départ de cette étude a été la remarque suivante : lorsque 30% d'individus possédent 60% de la masse totale, que vaut-il mieux dire?

Que ce groupe a une masse supérieure de 0,3 à son effectif (aspect Gini) ou qu'il posséde 2 fois plus en masse qu'en effectif?

C'est en adoptant le deuxième point de vue que j'ai abouti à cette nouvelle méthode d'analyse de la concentration d'une série : la méthode *mse* (pour rapport masse sur effectif : $\frac{\text{masse}}{\text{effectif}}$).

Je vais alors montrer que l'analyse de la concentration d'une série à l'aide du coefficient *g* de Gini, qui est la valeur moyenne des différences masse moins effectif peut être avantageusement remplacée par une analyse de deux indicateurs de concentration (c_{50} pour la concentration globale, c_{10} pour la concentration finale) qui eux, ne sont pas des différences, ni des moyennes, mais des rapports masse sur effectif (*mse*) : c_{50}, qui varie entre 1 et 2, est le rapport masse possédée par les 50% derniers indvidus sur 0,5 (leur effectif) et c_{10}, qui varie entre 1 et 10 est le rapport masse possédée par les 10% derniers individus sur 0,1 (leur effectif).

Il est beaucoup plus simple de calculer l'ensemble de ces deux indicateurs, que d'obtenir le seul coefficient de Gini ; et pour une série décilée, c_{50} et c_{10} se calculent de tête!

Notons tout de suite que la répartition égalitaire caractérisée par g =0 lorsqu'on utilise Gini, est aussi caractérisée par $c_{50} = 1$ ou $c_{10} = 1$ (les deux égalités sont équivalentes!)

Ces deux indicateurs permettent des conclusions plus pertinentes que la seule considération du coefficient de Gini ; ils se complétent l'un l'autre, un peu comme l'écart-type d'une série vient compléter sa valeur moyenne.

Par exemple, pour une série où $g = 0,08$, on va conclure à une répartion presque égalitaire alors qu'on peut avoir $c_{50} = 1,08$ et $c_{10} = 1,75$: les 50% derniers individus , qui possédent 50,4% de la masse totale, ont donc presque autant que les 50% premiers individus, qui eux possédent 49,6% de la masse totale :, donc au niveau global il y a répartition presque égalitaire.

Mais c_{10} permet de remarquer que les 10% derniers individus possédent 1,75 fois plus en masse qu'en effectif, et donc, pour ces 10% derniers individus il n'y a pas répartition égalitaire, mais un peu de concentration, ce que ne détecte évidemment pas le coefficient de Gini.

Table des matières

1 Introduction

On considère une série d'individus dont on étudie un caractère quantitatif prenant les valeurs $x_1, x_2, \ldots, x_p$ avec $p \geq 1$, $x_1 > 0$, $x_i < x_{i+1}$ et on s'intéresse au probléme suivant : existe-t-il un groupe d'individus possédant une part de la masse nettement supérieure à son poids démographique, par exemple un groupe ayant 60% de la masse totale (la masse salariale d'une entreprise par exemple) alors que son effectif n'est que de 30% de l'effectif total. Dans ce cas peu d'individus (30%) ont beaucoup (60% de la masse salariale) ou beaucoup d'individus (70%) ont peu (40% de la masse salariale) : on parle alors de ***concentration***.

Une méthode habituelle pour répondre à cette question est la méthode de Lorentz-Gini qui consiste à faire une courbe (la courbe de Lorentz ou de Gini ou de concentration) et/ou à calculer le coefficient de Gini.

Dans un premier temps je décrirai en détail cette méthode en la complétant par quelques résultats qui n'apparaissent pas dans la littérature sur ce sujet.

1) Il est impossible que le coefficient de Gini, noté g, prenne la valeur 1.

2) On a toujours $g \leq \dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$.

3) Pour toute série décilée $g < 0,9$.

On constate que ***ce coefficient de Gini, a deux inconvénients majeurs*** : il est lourd à calculer et il n'apporte pas d'éclairage précis sur la série. En effet lorqu'on a trouvé $g = 0,66$ (série des patrimoines 1986 par exemple) et que l'on conclut à forte concentration (alors que g n'est pas si éloigné que cela de la valeur centrale 0,5) qu'apporte-t-on de plus par rapport à la répartition de la série? En tout cas il n'y a pas de lien quantitatif précis entre ce coefficient et les groupes ayant beaucoup plus en masse qu'en effectif.

L'analyse théorique de la courbe de Lorentz (caractéristique de la série à deux facteurs d'échelle prés) permet de mettre en évidence un nouveau coefficient égal à l'écart maximum entre masse et effectif. Ce nouveau coefficient, noté e, est considérablement plus facile à calculer que g et s'analyse exactement de la même façon que g : si $e = 0$ il y a répartition égalitaire, si $e \simeq 1$ beaucoup ont peu ou peu ont beaucoup. ***Compte tenu de sa simplicité de calcul il peut remplacer avantageusement*** *g.* ***Cependant, tout comme*** *g,* ***ce coefficient ne permet pas une analyse précise de la série (en terme de concentration), cela parce que ces deux coefficients reposent sur la même idée : pour comparer deux nombres, on fait leur différence.***

Or si l'on considére un groupe ayant 60% de la masse totale pour un effectif de 30% vaut-il mieux dire que ce groupe a une masse supérieure de 0,3 à son effectif (60%-30%) ou qu'il a 2 fois plus en masse qu'en effectif $\dfrac{60\%}{30\%}$?

En général c'est plutôt des rapports que l'on considére : par exemple le rendement d'un appareil éléctrique est le rapport entre la puissance de sortie et la puissance d'entrée (entre 0 et 1) et non la différence entre ces deux quantités. De même pour une action, le PER est le rapport entre le cours de l'action et le bénèfice net par action (rapport qui lui peut être beaucoup plus grand que 1) et non la différence de ces 2 quantités. On peut aussi citer la notion d'élasticité qui est un rapport entre deux variations relatives et qui mesure la sensibilité de la demande à la variation de prix.

Une autre approche possible pour quantifier la notion de concentration est donc de considérer non pas les différences effectif moins masse mais les ***rapports masse sur effectif*** *(mse)*.

On s'apercoit alors que la concentration d'une série est une notion à plusieurs composantes *:* ***c'est un vecteur***. Deux composantes principales peuvent être mises en évidence (toutes les séries ayant, respectivement, ces mêmes composantes ont des courbes de Lorentz qui passent par quatre mêmes points) :

1ére composante : le *mse* des 50% derniers individus (compris entre 1 et 2), que l'on peut considérer comme un indicateur de concentration globale.

2iéme composante : le *mse* des 10% derniers individus (compris entre 1 et 10), que l'on peut considérer comme un indicateur de concentration finale.

Ces deux composantes sont très faciles à calculer et ont chacune une interprétation très simple et très concréte. En fait de part sa nature intégrale le coefficient de Gini est une valeur moyenne, une sorte d'indicateur global : d'ailleurs, à une transformation affine près ce coefficient est pratiquement égal à la première composante ci-dessus. Au prix de calculs compliqués on perd ainsi de l'information : par exemple, la longueur d'un vecteur ou la moyenne de ses composantes ne permet pas de connaître exactement ce vecteur.

La considération ***SIMULTANEE*** de ces deux composantes permet de définir une nouvelle méthode simple et efficace pour étudier la concentration d'une série. Notons que pour que la répartition soit égalitaire il faut et il suffit que l'une des composantes soit égale à sa valeur minimum 1 (auquel cas d'ailleurs l'autre composante est égale à 1)

.

Le lecteur désirant mettre en pratique tout de suite cette ***nouvelle méthode*** *mse* peut se reporter directement au chapitre 9 où il trouvera toutes les justifications théoriques nécéssaires, ainsi que les modalités d'application pratique illustrées par de nombreux exemples. Il constatera immédiatement la simplicité et l'efficacité de cette méthode, aussi bien pour l'étude d'une série que pour la comparaison de deux séries ou l'évolution temporelle d'une série. Par exemple pour la série des patrimoines 1986 la masse des 50% derniers individus est égale à 94% donc leur rapport masse sur effectif est 1,88 proche de la valeur maximale 2 : les 50% derniers individus ont presque deux fois plus en masse qu'en effectif, il y a donc ***concentration globale extrême*** ; par ailleurs la masse des 10% derniers individus est 53,8% donc leur rapport masse sur effectif est 5,38 (ils ont 5,38 fois plus en masse qu'en effectif) : ce rapport pouvant varier entre 1 et 10, on peut considérer que ***la concentration finale est moyenne***.

Objectivement cette méthode nécessite moins de calculs et apporte des conclusions plus précises que la méthode du coefficient de Gini (qui ici est égal à 0,66).

Quant aux justications mathématiques présentes dans cet ouvrage, pour la plupart un bon niveau de terminale scientifique est suffisant.

2 Rapports masse sur effectif, repartition egalitaire, concentration

2.1 Notations et hypothèses

On considère une série d'individus dont on étudie un caractère quantitatif prenant les valeurs $x_1, x_2, \ldots, x_p$ avec $p \geq 1$, $x_1 > 0$, $x_i < x_{i+1}$. Dans un cas continu les milieux des classes seront les x_i et on supposera que tous les individus d'une classe ont comme valeur du caractère le milieu de la classe. Les effectifs correspondants à ces valeurs du caractère seront notés $n_1, n_2, \ldots, n_p$ ***tous strictement positifs***.

On posera $n = \sum_{i=1}^{p} n_i$ l'effectif total et $m = \sum_{i=1}^{p} n_i x_i$ la masse totale (masse salariale si le caractère est le salaire, chiffre d'affaires d'une entreprise si l'individu est un client et le caractère le montant de l'achat). Enfin $\bar{x} = \frac{m}{n}$ désignera la moyenne de la série ($\bar{x} > 0$).

On pourra parfois envisager le cas $x_1 = 0$ mais pour $p \geq 2$ (ce qui assure encore $\bar{x} > 0$) ; cela sera alors explicitement dit.

.

2.2 Quelques définitions et résultats simples

Avant de quantifier, mesurer la concentration d'une série, il faut d'abord préciser cette notion : la plupart des ouvrages sur la question se contentent de dire qu'il y a forte concentration si peu ont beaucoup ou beaucoup ont peu, par exemple 30% ont 60% de la masse. Mais comment quantifier cet aspect : en considérant 0,6-0,3=0,3 ou 60/30=2?

Le rapport 60/30 me parait plus parlant : ce groupe a deux fois plus en masse qu'en effectif. Pour cette raison je vais développer la notion de rapport masse sur effectif.

2.2.1 Définition

♯ Etant donné un groupe G d'individus on note $mse(G)$ le rapport entre la masse (en %)
♯ possédée par ce groupe et l'effectif (en %) de ce groupe.

Il est évident qu'à effectif égal les groupes G ayant le plus fort *mse* seront les groupes d'individus ayant la plus forte valeur du caractère : précisons cela.

2.2.2 Propriété

1) ‖ Si G est le groupe constitué par tous les individus de la série on a $mse(G) = 1$
2) ‖ Quelque soit le groupe G d'effectif non nul on a $mse(G) = \frac{\bar{x}_G}{\bar{x}}$
   ‖ avec $\bar{x}_G$ moyenne du caractère pour les individus du groupe G
3) ‖ le *mse* du groupe (ou tout sous-groupe) des n_k individus ayant la valeur x_k du caractère
   ‖ est égal à $\frac{x_k}{\bar{x}}$

Preuve :

1) Evident.

2) Supposons que G soit constitué de k individus ayant les valeurs du caractère $x_{i_1}, x_{i_2} \ldots . x_{i_k}$, alors

$$mse(G) = \frac{\frac{x_{i_1} + x_{i_2} + \ldots . x_{i_k}}{m}}{\frac{k}{n}} = \frac{\frac{x_{i_1} + x_{i_2} + \ldots . x_{i_k}}{k}}{\frac{m}{n}} = \frac{\bar{x}_G}{\bar{x}}$$

3) Evident d'aprés le point 2.

2.2.3 Propriété

1) ‖ Si $p = 1$ quelque soit le groupe G on a $mse(G) = 1$
2) ‖ Si $p \geq 2$ alors $x_1 < \bar{x} < x_p$
   ‖ et quelque soit le groupe G, $mse(G) \in \left[\frac{x_1}{\bar{x}}; \frac{x_p}{\bar{x}} \right] \subset \left] \frac{x_1}{x_p}; \frac{x_p}{x_1} \right[$
3) ‖ Les groupes ayant les plus forts mse ($\frac{x_p}{\bar{x}}$) sont les sous-groupes des n_p derniers individus
   ‖ (ceux ayant la valeur x_p du caractère), en particulier le proupe des n_p premiers individus.
4) ‖ Les groupes ayant les plus faibles mse ($\frac{x_1}{\bar{x}}$) sont les sous-groupes des n_1 premiers
   ‖ individus (ceux ayant la valeur x_1 du caractère), en particulier le groupe des n_1
   ‖ premiers individus.

Preuve :

1) Si $p = 1$ tous les individus ont comme valeur du caractère x_1, donc pour tout groupe G sa moyenne $\bar{x}_G$ est égale à $x_1 = \bar{x}$ et donc $mse(G) = 1$.
2)Si $p \geq 2$ les hypothèses faites au 2.1 entraînent $n_i x_i > n_i x_1$ cela pour i variant de 2 à p donc $\sum_{i=1}^{p} n_i x_i > (\sum_{i=1}^{p} n_i) x_1 = n x_1$ d'où $\bar{x} > x_1$, de même $\sum_{i=1}^{p} n_i x_i < (\sum_{i=1}^{p} n_i) x_p$ et $\bar{x} < x_p$: finalement $x_1 < \bar{x} < x_p$.
Pour un groupe G quelconque constitué de k_i individus ($k_i \geq 0$) ayant la valeur x_i (cela pour i variant de 1 à p), on aura cette fois $k_i x_i \geq k_i x_1$ et finalement $x_1 \leq \bar{x}_G \leq x_p$.
En divisant par $\bar{x}$ on obtient $mse(G) \in \left[\frac{x_1}{\bar{x}}; \frac{x_p}{\bar{x}} \right]$.

3) $mse(G) = \frac{x_p}{\bar{x}} \Leftrightarrow \bar{x}_G = x_p$ qui n'est possible que si tous les individus du groupe G ont x_p comme valeur du caractère (résultat classique sur moyenne arithmétique)

4) Même raisonnement qu'en 2.

Précisons maintenant la notion de répartition égalitaire d'une série:

2.2.4 Définition

♯ Une répartition égalitaire est une répartition pour laquelle tout groupe d'individus a autant
♯ en masse qu'en effectif, c'est-à-dire tout groupe d'individus a un mse égal à 1.

2.2.5 Propriété

‖ Une répartition est égalitaire si et seulement si tous les individus ont la même valeur
‖ du caractère c'est-à-dire $p = 1$ ou autrement dit l'étendue de la série est nulle.

Preuve :

Si la répartition est égalitaire alors pour tout groupe G on a $mse(G) = 1,$ et d'aprés 2.2.2 $\forall k \in \{1;2;3;\ldots;p\}$ on a $\frac{x_k}{\bar{x}} = 1$ donc $p = 1,$ car si $p \geq 2$ on aurait $x_1 < x_2 <\ldots< x_p$ ce qui est en contradiction avec le fait que tous les x_i doivent être égaux (à $\bar{x}$). Réciproquement si $p = 1$ tout groupe G a un *mse* égal à 1 (voir 1 de 2.2.3) et la répartition est bien égalitaire.

2.2.6 Définition

♯ Il y a concentration signifie qu'il n'y a pas répartition égalitaire.

2.2.7 Propriété

1) ‖ Il y a concentration si et seulement si il existe un groupe dont le *mse* est différent de 1.
2) ‖ Il y a concentration si et seulement si $p \geq 2$ (c'est à dire l'étendue $x_p - x_1$ est non nulle).

Preuve :

1) et 2) sont des conséquences immédiates de 2.2.4, 2.2.5, 2.2.6.

Précisons un lien qualitatif entre l'étendue $x_p - x_1$ de la série et la notion de concentration : si x_1 est peu différent de x_p (cas de faible étendue) alors pour tout groupe G, $mse(G)$ est peu différent de 1 puisque $\bar{x}$ et $\bar{x}_G$ sont compris entre x_1 et x_p et donc eux aussi sont peu différents. Réciproquement si pour tout groupe G, $mse(G) \simeq 1$ on a $\frac{x_1}{\bar{x}} \simeq 1$ et $\frac{x_p}{\bar{x}} \simeq 1$ donc x_1 et x_p sont peu différents. D'où la propriété suivante :

2.2.8 Propriété

‖ $x_p \simeq x_1$ équivaut au fait que pour tout groupe G, $mse(G)$ est peu différent de 1, ce que l'on
‖ peut qualifier de répartition ‹‹presque›› égalitaire : ***l'aspect concentration*** ne peut
‖ vraiment exister que s'il y a forte étendue, c'est à dire si x_1 et x_p sont trés différents.

Remarque 1 :
Evidemment ce résultat n'est pas une découverte ; on verra au chapitre 9 qu'en fait si $\frac{x_p}{x_1} < 1,25$ alors pour tout groupe G d'effectif supérieur ou égal à 10% on a $mse(G) \leq 1,22$: on peut estimer que l'aspect concentration n'existe pratiquement pas.

Remarque 2 :
Pour le moment la notion de concentration d'une série reste une notion qualitative : il y a ou il n'y a pas concentration. La méthode de Gini consiste justement à quantifier par un seul chiffre cette notion de concentration, chiffre égal à zéro lorqu'il y a répartition égalitaire d'où l'expression concentration nulle pour désigner un cas de répartition égalitaire. En fait je montrerai plus loin que quantifier la concentration d'une série par un seul chiffre est trop réducteur.

Remarque 3 :
On pourrait penser que dans le cas où x_1 est trés différent de x_p, mais avec un effectif trés supérieur aux autres (par exemple n_k) tout se passe comme si presque tous les individus avaient la même valeur du caractère x_k et donc qu'il y a ‹‹presque›› répartition égalitaire : en fait cela est faux puique l'on vient de justifier qu'une telle situation exige $x_p \simeq x_1$. Vérifions le sur un exemple :

x_i	n_i	$n_i x_i$
1	2	2
2	3	6
8	90	720
16	10	160
	$n = 105$	m =888

On a bien x_1 trés différent de x_4, un effectif trés supérieur aux autres et pourtant le groupe des 10 derniers individus (presque 10% de l'effectif total) a un *mse*=$\frac{x_4}{\bar{x}} = 16 \times \frac{105}{888} \simeq 1,89$ assez supérieur à 1 : ce groupe a presque deux fois plus en masse qu'en effectif. On ne peut pas dire que la répartition soit presque égalitaire.

3 Méthode de Gini Lorentz

3.1 Description de la méthode

On peut la décomposer en 4 phases : calcul des effectifs et masses en cumulés croissants, construction de la courbe de Lorentz, calcul du coefficient de Gini , analyse.

3.1.1 Phase1 Calcul des effectifs et masses en cumulés croissants

On calcule les effectifs α_k (en %) et masses β_k (en %) en cumulés croissants :

pour $k \in \{1;2;\ldots;p\}$ $\quad \boxed{\alpha_k = \dfrac{\sum_{i=1}^{k} n_i}{n} \text{ et } \beta_k = \dfrac{\sum_{i=1}^{k} n_i x_i}{m}}$

Par commodité on posera $\alpha_0 = \beta_0 = 0$. Notons que $\alpha_p = \beta_p = 1$ et bien sûr $0 < \beta_1 < \beta_2 < \ldots < \beta_p$ ainsi que $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_p$.

Interprétation : pour $k \geq 1$ α_k est le pourcentage des individus ayant une valeur du caractère inférieure ou égale à x_k : ces individus possédent une fraction de la masse totale égale à β_k.

On peut aussi dire que β_k est la masse (en %) possédée par les α_k premiers individus (en %), ceux-ci étant classés par valeur croissante du caractère.

3.1.2 Phase2 Construction de la courbe de Lorentz

On reporte les points $M_k(\alpha_k, \beta_k)$ pour $k \in \{0;1;\ldots;p\}$ sur un graphique, et on cherche à les relier par une courbe qui peut être considérée comme la représentation graphique d'une fonction f. Cette fonction va alors vérifier $f(\alpha_k) = \beta_k$, c'est-à-dire $f(\alpha_k)$ sera la masse (en %) possédée par les α_k ***premiers*** individus (en %) ; l'idéal serait donc de choisir f de façon que pour α quelconque de la forme $\frac{k}{n}$ (avec k entier entre 0 et n) on ait ***encore*** $f(\alpha)$ égal à la masse possédée par les α ***premiers*** individus (en %).

On verra au 4.1 et 4.2 que la ligne brisée constituée par les segments $[M_k M_{k+1}]$ correspond justement à une fonction f vérifiant cette propriété.

On appelera donc courbe de Lorentz la ligne brisée constituée des segments $[M_k M_{k+1}]$ ***et on la notera*** C_{pr} (*pr* comme premier)

Voici un exemple :

On a toujours $M_p = A$ avec $A(1,1)$ et on note B le point de coordonnées $(1,0)$

Avant d'aborder la phase 3 précisons les propriétés de cette courbe C_{pr} :

3.1.3 Propriété de la courbe de Lorentz

1) ‖ Si $p = 1$ la courbe C_{pr} est le segment $[OA]$
SI $p \geq 2$ (et même si $x_1 = 0$) ALORS
2) ‖ $\frac{\beta_1}{\alpha_1} < \frac{\beta_2}{\alpha_2} <\ldots< \frac{\beta_{p-1}}{\alpha_{p-1}} < 1$
3) ‖ $\beta_k < \alpha_k$ pour tout $k \in \{1;2;\ldots;p-1\}$
4) ‖ La pente de la droite (OM_k) est inférieure à la pente de la droite (OM_{k+1})
5) ‖ La courbe C_{pr} est strictement en dessous du segment $[OA]$ (sauf pour $M_0 = O$ et $M_p = A$)
6) ‖ Les pentes des segments $[M_kM_{k+1}]$ sont respectivement égales à $\frac{x_{k+1}}{\bar{x}}$: elles forment
‖ une suite strictement croissante lorsque k varie de 0 à $p-1$.
7) ‖ C_{pr} admet uniquement $p-1$ points anguleux : les points $M_1, M_2, \ldots, M_{p-1}$
8) ‖ C_{pr} est la représentation graphique d'une fonction convexe et affine par intervalles.
9) ‖ Les points M_k s'éloignent d'abord de $[OA]$ puis s'en rapprochent : c'est la traduction
‖ géométrique du fait que la suite $\alpha_k - \beta_k$ est d'abord croissante puis décroissante.

Preuve :

1) Si $p = 1$ il n'y a évidemment que deux points M_k : les points $M_0 = O$ et $M_1 = A$

2) et 3) Voir annexe 1.

4) à 9) Voir annexe 2.

3.1.4 Phase 3 Calcul du coefficient de Gini

On calcule l'aire a de la région délimitée par C_{pr} et le segment $[OA]$ (ce qui peut se faire en comptant les carreaux: voir réference 3 et 5) ou en appliquant la formule :

$$a = \frac{1}{2} - \frac{1}{2n}\sum_{i=0}^{p-1} n_{i+1}(\beta_i + \beta_{i+1})$$

Cette formule s'obtient en remarquant que la région située sous C_{pr} se décompose en p trapézes de hauteurs $\frac{n_{i+1}}{n}$ et de bases β_i et β_{i+1}. On verra une autre preuve à l'aide d'un calcul intégral au chapitre 4.
Et enfin on calcule le coefficient ou indice de concentration de Gini g :

$$g = \frac{a}{\text{aire du triangle (OAB)}} = 2a$$

soit

$$g = 1 - \sum_{i=0}^{p-1} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1}) = 1 - \sum_{i=0}^{p-1} f_{i+1}(\beta_i + \beta_{i+1})$$

$f_i = \frac{n_i}{n}$ étant la fréquence d'apparition du ième caractère.

Cas des séries décilées *:*

Une série décilée est une série où les n individus sont répartis en 10 classes de mêmes effectifs selon le principe suivant : les individus étant classés par ordre de caractère croissant, la première classe est constituée des 10% premiers individus, la 2ième classe des 10% individus suivants.....
Notons q_i la masse de la ième classe (on supposera ici que $q_1 < q_2 < \ldots < q_{10}$) : en prenant $p = 10$, $n_i = \frac{n}{10}$ et $x_i = \frac{10}{n}q_i$ pour $i = 1,2,\ldots,10$, la série décilée est alors décrite par le modèle du 2.1 (cela revient à supposer que tous les individus d'une classe ont la même valeur du caractère).
On a alors $m = \sum_{i=1}^{10} q_i$, $\alpha_k = \frac{k}{10}$, $\beta_k = \frac{\sum_{i=1}^{k} n_i x_i}{m} = \frac{\sum_{i=1}^{k} q_i}{m}$ et

$$g = \frac{9}{10} - \frac{\beta_1 + \beta_2 + \ldots + \beta_9}{5}$$

(voir le détail du calcul dans le cadre de la preuve du 3 du 3.2.3)
On notera que pour ce modèle la suite $n_i x_i = q_i$ est strictement croissante ce qui n'est pas le cas en général.

L'interprétation de ce coefficient de Gini (voir phase 4) repose sur la propriété suivante :

3.1.5 Propriétés du coefficient de Gini

1) $g \in [0;1]$ et plus g est grand plus C_{pr} est ‹‹éloignée›› de $[OA]$: g est une mesure de la proximité de C_{pr} à $[OA]$
2) $g = 0 \Leftrightarrow C_{pr}$ est confondue avec $[OA] \Leftrightarrow p = 1 \Leftrightarrow$ la répartition est égalitaire
3) $g > 0 \Leftrightarrow$ il y a concentration
4) g proche de 1 $\Leftrightarrow$ beaucoup (les premiers individus) ont peu
   $\Leftrightarrow$ peu (les derniers individus) ont beaucoup.

On démontrera au paragraphe suivant qu'il est impossible que g ***prenne la valeur 1***,
et on verra en annexe 4 l'influence sur g d'un ajout d'une même quantité à toutes les valeurs du caractère.

Preuve :

1) Evident

2) $g = 0 \Leftrightarrow a = 0$ ce qui donne la première équivalence. Pour les autres : si C_{pr} est confondue avec $[OA]$ alors C_{pr} n'a pas de point anguleux et donc $p = 1$ d'aprés le 7 de 3.1.3 et donc la répartition est égalitaire. Réciproquement s'il y a répartition égalitaire alors $p = 1$ et C_{pr}=$[OA]$

3) Evident puisque concentration signifie qu'il n'y a pas répartition égalitaire (2.2.6).

4) Si g est proche de 1 alors C_{pr} est proche de $[OB] \cup [BA]$ et donc il existe un point M_{k_0} tel que $\alpha_{k_0} \simeq 1$ et $\beta_{k_0} \simeq 0$, c'est-à-dire beaucoup (les α_{k_0} premiers individus) ont peu (β_{k_0}). Réciproquement s'il existe k_0 tel que $\alpha_{k_0} \simeq 1$ et $\beta_{k_0} \simeq 0$ le point M_{k_0} est proche de B et compte tenu de la position relative des points M_k (voir 3.1.3) la courbe C_{pr} est entiérement située dans la zone hachurée de la figure ci dessous et donc C_{pr} est effectivement proche de $[OB] \cup [BA]$, c'est-à-dire g proche de 1.

3.1.6 Phase 4 Analyse du coefficient de Gini

La propriété précédente permet d'affirmer que le coefficient g est effectivement un indicateur de concentration : plus g augmente, plus g est proche de 1, plus C_{pr} se rapproche de $[OB] \cup [OA]$ et plus il existe un point $M_{k_0}(\alpha_{k_0}, \beta_{k_0})$ proche de $B(1, 0)$ c'est-à-dire beaucoup (α_{k_0})ont peu (β_{k_0}) ou peu ($1 - \alpha_{k_0}$) ont beaucoup ($1 - \beta_{k_0}$).

Mais dire g mesure ***la*** ‹‹concentration›› de la série et donc ***la*** ‹‹concentration›› d'une série varie entre 0 (cas de concentration nulle ou répartition égalitaire) à 1 (concentration maximum) me parait discutable. Considérons par exemple la série des patrimoines de 1986 (voir 6.2) où $g = 0,66$: tout le monde en déduit alors qu'il y a forte concentration (bien que g ne soit pas si éloigné que cela de la valeur centrale 0,5 !) mais la seule connaissance de g ne permet pas de savoir d'où provient cette concentration. D'ailleurs on rajoute souvent pour en confirmer l'importance, le commentaire suivant : ‹‹en effet les 10% derniers individus ont 53,8% de la masse››. En fait g n'a aucun rapport direct avec les rapports masse sur effectif et n'a pas de signification précise en terme de concentration. On a simplement quantifié brutalement (par un seul chiffre) la notion de concentration, notion qui jusqu'à présent n'est définie que comme étant le contraire de la répartition égalitaire.

On verra au chapitre 8 que la notion de concentration ne peut être résumée par un seul chiffre.

A ce niveau le lecteur, s'il en éprouve le besoin, peut se reporter au chapitre 6 pour y trouver des exemples mais il devra ignorer les passages relatifs au coefficient e qui sera défini au chapitre 5.

3.2 Sur les valeurs extrêmes de g : 0 et 1

Au préalable une propriété qui servira plusieurs fois :

3.2.1 Propriété

1) ‖Si $M(\alpha,\beta)$ est un point situé à l'intérieur du triangle OAB alors
‖ l'aire du triangle OMA est $\frac{1}{2}(\alpha-\beta)$
2) ‖ Si $p=2$ le coefficient de Gini est $\alpha_1-\beta_1$
3) ‖ Si $M_1(\alpha_1,\beta_1)$ est un point quelconque situé à l'intérieur du triangle OAB alors
‖ il existe des séries dont la courbe de Lorentz est la ligne brisée $[OM_1]\cup[M_1A]$

Preuve :

1) Soit I le projeté parallélement à (OB) de $M(\alpha,\beta)$ sur $[OA]$ et H le projeté orthogonal de M sur $[OA]$:

on a évidemment $IM=\alpha-\beta$, d'où $\sin(\frac{\pi}{4})=\frac{HM}{IM}$ et $HM=\frac{\sqrt{2}}{2}(\alpha-\beta)$ et donc l'aire du triangle OMA est
$\frac{1}{2}OA\times HM=\frac{1}{2}\sqrt{2}\,\frac{\sqrt{2}}{2}(\alpha-\beta)=\frac{1}{2}(\alpha-\beta)$.

2) Si $p=2$ la courbe de Lorentz se réduit à la ligne brisée $[OM_1]\cup[M_1A]$ et d'aprés 3.1.4 g est le double de l'aire du triangle OM_1A soit $2\frac{1}{2}(\alpha_1-\beta_1)=\alpha_1-\beta_1$. Bien entendu on obtient le même résultat en utilisant la formule
$g=1-\sum_{i=0}^{p-1}\frac{n_{i+1}}{n_i}(\beta_i+\beta_{i+1})$ du 3.1.4 : dans ce cas $p=2$, et puisque $\beta_0=0$, $\beta_2=1$ il vient
$g=1-(\frac{n_1}{n}+\frac{n_2}{n})\beta_1-\frac{n_2}{n}=\frac{n_1}{n}-\beta_1=\alpha_1-\beta_1$.

3) Il faut trouver (α_1 et β_1 étant donnés et vérifiant $0<\beta_1<\alpha_1<1$) les quatres nombres $n_1,\ n_2, x_1, x_2$ vérifiant les conditions :
$0<n_1,\ 0<n_2,\ 0<x_1<x_2, \alpha_1=\frac{n_1}{n}$ et $\beta_1=\frac{n_1x_1}{n_1x_1+n_2x_2}$.
La condition $\alpha_1=\frac{n_1}{n}$ détermine la fréquence $\frac{n_1}{n}$ et donc l'autre fréquence $\frac{n_2}{n}$;
Comme $\frac{n_1x_1}{n_1x_1+n_2x_2}=\frac{\alpha_1x_1}{\alpha_1x_1+(1-\alpha_1)x_2}=\frac{\alpha_1}{\alpha_1+(1-\alpha_1)\frac{x_2}{x_1}}$, la dernière condition donne le rapport
$\frac{x_2}{x_1}=\frac{\alpha_1}{\beta_1}\frac{1-\beta_1}{1-\alpha_1}$ rapport qui est bien supérieur à 1 puisque $\frac{\alpha_1}{\beta_1}>1$ et $\frac{1-\beta_1}{1-\alpha_1}>1$.
Finalement les fréquences sont déterminées de façon unique et donc les effectifs sont déterminés à une contante multiplicative près ; de même les deux valeurs du caractère sont déterminées à une constante multiplicative près.

3.2.2 Coefficient de Gini et valeur 0

On a vu au 3.1.5 que $g = 0$ équivaut à répartition égalitaire, mais il peut exister des séries (particulières certes) où g peut être trés proche de 0 avec la présence d'un groupe d'individus ayant nettement plus en masse qu'en effectif ce qui ne correspond pas à une répartition presque égalitaire (tous les *mse* sont peu différents de 1, voir 2.2.8).

Voici un exemple illustrant cela : on prend $p = 2$, $x_2 = 2x_1$, $n_1 = 1000$, $n_2 = 100$. D'aprés 3.2.1 on a $g = \alpha_1 - \beta_1 = \frac{n_1}{n_1 + n_2} - \frac{n_1 x_1}{n_1 x_1 + n_2 x_2} = \frac{1000}{1100} - \frac{1000}{1200} \simeq 0,0757$, qui est donc très faible. Pour autant, a-t-on le droit de dire que la ‹‹concentration›› est pratiquement nulle, c'est à dire que la répartition est presque égalitaire?

En fait le groupe des 100 derniers individus (9% de l'effectif) a un *mse* égal à $\frac{x_2}{\bar{x}} = \frac{x_2}{\frac{1000x_1 + 100x_2}{1100}} = \frac{2200}{1200} \simeq 1,83$: ce groupe a presque deux fois plus en masse qu'en effectif. Quant au groupe des 10% derniers individus il a um *mse* égal à $10\frac{10x_1 + 100x_2}{1000x_1 + 100x_2} = 1,75$ ce qui est encore trés supérieur à 1.

Le fait que g ***soit trés faible n'entraîne donc pas qu'il y a une répartition presque égalitaire*** (au sens du 2.2.8). Cela est dû au fait que le coefficient de Gini est une valeur moyenne (voir 4.5) et donc il y a perte d'information ; d'ailleurs g prend surtout en compte ce qui se passe au niveau des 50% derniers individus puisque d'aprés 4.6 une valeur approchée de g est $\frac{4}{3}$ ×masse des 50% derniers $-\frac{2}{3}$.

Notons cependant que le *mse* des 10% derniers individus peut en fait varier de 1 à 10 (voir le 2 de 7.6) et que 1,75 est tout de même plus proche de 1 (valeur minimum) que de 10.

3.2.3 Coefficient de Gini et valeur 1

S'il est certain que g ***est situé dans l'intervalle*** $[0;1]$ ***contrairement à ce qui est parfois dit*** g ***ne peut pas prendre la valeur 1***.

La propriété suivante précise cet aspect

Propriété

(Rappel : cf 2.1 on a toujours $x_1 > 0$, sauf exceptions explicitement signalées)

1) ‖ On a toujours $g \leq 1 - \frac{n_p}{n} < 1$ (vrai même si $x_1 = 0$ et $p \geq 2$)
   ‖ Sur l'ensemble de toutes les séries 1 est un majorant (non atteint) de g.
2) ‖ On a toujours $g \leq \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}} < 1$
   ‖ $\frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ est un majorant de g pour toutes les séries à x_1 et x_p donnés.
   ‖ Ce majorant est atteint si et seulement si $p = 2$ et $\frac{n_1}{n} = \frac{\sqrt{x_2}}{\sqrt{x_1} + \sqrt{x_2}}$.
3) ‖ Dans le cas d'une série décilée $g < 0,9$.
4) ‖ Sur l'ensemble de toutes les séries g décrit tout l'intervalle $[0;1[$.

preuve :

1) D'aprés 3.1.4 on a $g = 1 - A$ avec $A = \sum_{i=0}^{p-1} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1})$, mais toutes les quantités étant positives ou nulles ($\beta_0 = 0$ et $\beta_1 = 0$ si $x_1 = 0$) A est supérieur ou égal au terme correspondant à $i = p - 1$ lequel est supérieur ou égal à $\frac{n_p}{n}\beta_p = \frac{n_p}{n}$ (β_p est toujours égal à 1 même si $x_1 = 0$ et $p \geq 2$) et donc $g \leq 1 - \frac{n_p}{n}$, quantité STRICTEMENT inférieure à 1 puisque $n_p > 0$. Remarquons que si $p = 1$ cette majoration donne $g \leq 0$ et on retrouve $g = 0$.

2) Voir annexe 3. Il est à noter que la deuxième inégalité exige $x_1 > 0$, hypothèse générale faite au 2.1.

3) Il a été écrit au 3.1.4 que dans le cas d'une série décilée $g = \frac{9}{10} - \frac{\beta_1 + \ldots\ldots + \beta_9}{5}$.

Justifions le : d'aprés 3.1.4 $g = 1 - \sum_{i=0}^{p-1} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1})$ et puisque $p = 10$, $\frac{n_i}{n} = 10\%$, $\beta_0 = 0$, $\beta_{10} = 1$, on obtient $g = 1 - \frac{1}{10}(2(\beta_1 + \ldots\ldots + \beta_9) + 1)$ ce qui donne le résultat ci-dessus.

D'aprés $\beta_1 < \beta_2 < \ldots < \beta_9$ il vient $\beta_1 + \beta_2 + \ldots + \beta_9 > 9\beta_1$ et donc $g < \frac{9}{10} - \frac{9}{5}\beta_1$. Et comme x_1 et n_1 sont strictements positifs on a $\beta_1 > 0$ d'où $g < \frac{9}{10}$.

Notons que le résultat $g \leq 1 - \frac{n_p}{n}$ obtenu au 1 ci-dessus permettait seulement d'écrire $g \leq \frac{9}{10}$.

4) La valeur $g = 0$ est obtenue pour toute série telle que $p = 1$.

Soit u quelconque dans $]0;1[$.

Il existe u' vérifiant $0 < u' < 1 - u$. En posant $\beta_1 = u'$ et $\alpha_1 = u + u'$, le point $M_1(\alpha_1, \beta_1)$ est à l'intérieur du triangleOAB (car $0 < \beta_1 < \alpha_1 < 1$) et donc d'après 3.2.1 il existe une série ayant comme courbe de Lorentz la ligne brisée $[OM_1] \cup [M_1A]$ et donc son coefficient de Gini est $g = \alpha_1 - \beta_1 = u$.

Remarque 1 :

On peut s'étonner, compte tenu de l'impossibilité que g prenne la valeur 1, que l'on puisse rencontrer parfois l'affirmation suivante : si un seul salarié perçoit toute la masse alors $g = 1$. En fait cette situation exige que tous les autres ne reçoivent rien ce qui correspond au modèle suivant : $p = 2$, $x_1 = 0$, $x_2 > 0$, $n_1 \neq 0$, $n_2 = 1$ donc $\alpha_1 = \frac{n-1}{n}$, $\beta_1 = 0$ et $\alpha_2 = \beta_2 = 1$. La courbe de Lorentz est donc :

D'aprés 3.2.1 $g = \alpha_1 - \beta_1 = \frac{n-1}{n} < 1$; de même si k salariés possédaient tout (avec chacun le même salaire) on aurait $g = \frac{n-k}{n} < 1$. Dans les 2 cas g n'est pas égal à 1, par contre $\lim_{n \to +\infty} g = 1$.

Remarque 2 :

A titre de curiosité montrons par un raisonnement géométrique que (même si $x_1 = 0$ et $p \geq 2$) g ne peut être égal à 1, c'est à dire que la courbe de Lorentz ne peut être confondue avec $[OB] \cup [BA]$. Si tel était le cas p serait supérieur ou égal à 2 (sinon C_{pr} est confondue avec $[OA]$) et M_{p-1} ne pourrait être sur $[OB[$ (sinon le segment $[M_{p-1}A]$, qui est une partie de la courbe de Lorentz, ne pourrait être inclu dans $[OB] \cup [BA]$) donc il serait sur $[BA[$ et le segment $[M_{p-1}A]$ serait vertical ce qui est contradictoire avec le fait que sa pente, $\frac{x_p}{\bar{x}}$, est un nombre fini.

Bien entendu, on peut aussi dire que C_{pr} est la représentation graphique d'une fonction (m_{pr}) et à ce titre elle ne peut avoir de partie verticale et donc elle ne peut être confondue avec $[OB] \cup [BA]$.

Remarque 3 :

Si on pose $r = \frac{x_p}{x_1}$ alors $g \leq \frac{\sqrt{r} - 1}{\sqrt{r} + 1}$; or $\frac{\sqrt{r} - 1}{\sqrt{r} + 1}$ est une expression qui est croissante avec r, donc par exemple, $\frac{x_p}{x_1} \leq 16$ entraîne que quelque soient les effectifs on a $g \leq \frac{\sqrt{16} - 1}{\sqrt{16} + 1} = 0,6$.

3.3 Sur les séries ayant la même courbe de Lorentz

Il s'agit de généraliser en fait le point 3 de la propriété 3.2.1 :

Propriété

1) ‖ Soit $\widehat{C}$ une ligne polygonale reliant les points $O = N_0$ et $A = N_q$ $(q \geq 2)$ et dont
‖ les $q-1$ points anguleux $N_1(u_1, v_1), N_2(u_2, v_2), \ldots, N_{q-1}(u_{q-1}, v_{q-1})$ vérifient
‖ a) pente de $[N_i N_{i+1}]$ < pente de $[N_{i+1} N_{i+2}]$ pour $i = 0, 1, \ldots, q-2$
‖ b) $0 \leq v_1$ et $0 < u_1 < u_2 < \ldots < u_{q-1} < 1$
‖ alors toutes les séries (vérifiant les hypothèses du 2.1) et dont la courbe de Lorentz
‖ est $\widehat{C}$ sont les séries prenant q valeurs $x_1, x_2, \ldots, x_q$ avec les effectifs $n_1, n_2, \ldots, n_q$
‖ tels que $n_i = (u_i - u_{i-1})k_1$ et $x_i = \frac{v_i - v_{i-1}}{u_i - u_{i-1}} k_2$ pour $i = 1, 2, ..., q$
‖ où k_1 et k_2 sont 2 constantes arbitraires strictement positives (k_1 sera l'effectif total et
‖ k_2 sera la valeur moyenne de la série)
2) ‖ Si deux séries $p, x_1, x_2, \ldots, x_p, n_1, n_2, \ldots, n_p$ et $q, x'_1, x'_2, ..., x'_q, n'_1, n'_2, \ldots, n'_q$
‖ ont la même courbe de Lorentz alors $p = q$ et il existe deux constantes k_1 et k_2
‖ telles que $n_i = k_1 n'_i$ (les fréquences sont respectivement les mêmes) et $x_i = k_2 x'_i$ pout tout i.

Preuve :

1) Soit une série $p, x_1, x_2, \ldots, x_p, n_1, n_2, \ldots, n_p$ vérifiant les hypothéses du 2.1 : si sa courbe de Lorentz est $\widehat{C}$ alors d'aprés 3.1.3 $p - 1 = q - 1$ (même nombre de points anguleux) soit $p = q$ et les points anguleux sont les mêmes d'où pour $i = 0, 1, \ldots, p$ on a $\alpha_i = u_i$ et $\beta_i = v_i$ et donc $n_i = n(\alpha_i - \alpha_{i-1}) = n(u_i - u_{i-1})$ et pente de $[M_{i-1}M_i]$ = pente de $[N_{i-1}N_i]$ donne (voir aussi 3.1.3) $x_i = \frac{v_i - v_{i-1}}{u_i - v_{i-1}} \bar{x}$.
Réciproquement soit une série prenant q valeurs $x_1, x_2, \ldots, x_q$ avec les effectifs $n_1, n_2, \ldots, n_q$ tels que $n_i = (u_i - u_{i-1})k_1$ et $x_i = \frac{v_i - v_{i-1}}{u_i - u_{i-1}} k_2$: notons que les hypothéses $k_1 > 0$, $k_2 > 0$ et celles faites sur les u_i et les pentes de $[N_i N_{i+1}]$ assurent que tous les n_i sont strictement positifs et que $x_1 = \frac{v_1}{u_1} k_2 \geq 0$ (> 0 si $v_1 > 0$) et $x_1 < x_2 < \ldots < x_q$.
Les hypothéses du 2.1 sont donc vérifiées et la courbe de Lorentz C_{pr} de cette série est une ligne polygonale dont les sommets sont (pour $i = 0$ à q) les points $M_i(\alpha_i, \beta_i)$.
L'effectif total de cette série, somme des n_i, est alors $n = (u_q - u_0)k_1 = k_1$ (puique $N_0 = O$ et $N_q = A$) et sa valeur moyenne est $\bar{x} = \frac{\sum_1^q n_i x_i}{n} = \frac{1}{n} \sum_1^n (u_i - u_{i-1}) k_1 \frac{v_i - v_{i-1}}{u_i - u_{i-1}} k_2 = \frac{k_1 k_2}{n} = k_2$.
De $\alpha_i = \frac{n_1 + n_2 + \ldots\ldots n_i}{n}$ on déduit alors $\alpha_i = \frac{u_i k_1}{k_1}$ soit $\underline{\alpha_i = u_i}$.
Par ailleurs, d'aprés 3.1.3, la pente de $[M_{i-1}M_i]$ est égale à $\frac{x_i}{\bar{x}}$, ce qui donne $\frac{\beta_i - \beta_{i-1}}{\alpha_i - \alpha_{i-1}} = \frac{x_i}{k_2}$ soit $\frac{\beta_i - \beta_{i-1}}{u_i - u_{i-1}} = \frac{v_i - v_{i-1}}{u_i - u_{i-1}}$ d'où $\beta_i - \beta_{i-1} = v_i - v_{i-1}$ et comme $\beta_0 = v_0 = 0$ on a $\underline{\beta_i = v_i}$. On a donc montré que $M_i = N_i$ pour tout i de 0 à q et donc $C_{pr} = \widehat{C}$.

2) Prenons pour $\widehat{C}$ la courbe de Lorentz de la deuxiéme série : le résultat 1 que l'on vient de démontrer permet de dire que $p = q$, $n_i = (\alpha'_i - \alpha'_{i-1})k_1 = k_1 n'_i$, $x_i = \frac{\beta'_i - \beta'_{i-1}}{\alpha'_i - \alpha'_{i-1}} k_2 = \frac{x'_i}{\bar{x'}} k_2$ soit $x_i = k_2 x'_i$ en intégrant le facteur $\frac{1}{\bar{x'}}$ dans la constante k_2. Le fait que les fréquences soient respectivement les mêmes résulte de $n_i = k_1 n'_i$ pour $i = 1, 2, \ldots, p$: en effet ceci entraîne que $n = k_1 n'$ et donc $\frac{n_i}{n} = \frac{n'_i}{n'}$.

Remarque 1 :
On notera que pour le résultat 1 on n'a pas supposé explicitement que les points N_i sont dans le triangle *OBA* (c'est à

dire $0 \leq v_i \leq u_i \leq 1$) : cela résulte des hypothéses faites puisque celles-ci entrainent que $\widehat{C}$ est une courbe de Lorentz.

Remarque 2 :

Si $q = 2$ les hypothèses faites pour le 1 s'écrivent $0 \leq v_1$, $0 < u_1 < 1$, $\frac{v_1}{u_1} < \frac{1 - v_1}{1 - u_1}$ et on peut alors vérifier qu'elles sont équivalentes à $0 \leq v_1 < u_1 < 1$.

4 Compléments sur la courbe de Lorentz et le coefficient de Gini.

4.1 Défintion

1) ♯ On appelle m_{pr} la fonction affine par intervalles dont la représentation graphique, notée C_{pr},
♯ est la courbe de Lorentz (ligne brisée reliant les points M_k, voir 3.1.2)
2) ♯ On appelle m_{dr} la fonction affine par intervalles dont la représentation graphique, notée C_{dr},
♯ est la symétrique par rapport au milieu de $[OA]$ de la courbe de Lorentz.

4.2 Propriété

Les individus étant classés par valeur croissante du caractère et α s'interprétant toujours comme un pourcentage d'individus :

1) ‖ m_{pr} est une fonction convexe sur $[0;1]$; $m_{pr}(\alpha) = \frac{x_1}{\bar{x}}\alpha$ pour $\alpha \in [0;\alpha_1]$
‖ et pour α de la forme $\frac{k}{n}$ avec $k \in \{0;1;2;\ldots;n\}$ $m_{pr}(\alpha)$ est la ***masse (en pourcentage)***
‖ ***possédée par les α premiers individus*** (en particulier $m_{pr}(\alpha_i) = \beta_i$).
2) ‖ $\forall \alpha \in [0;1]$ on a $m_{dr}(\alpha) = 1 - m_{pr}(1-\alpha)$.
3) ‖ m_{dr} est une fonction concave sur $[0;1]$; $m_{dr}(\alpha) = \frac{x_p}{\bar{x}}\alpha$ pour $\alpha \in [0;1-\alpha_{p-1}]$
‖ et pour α de la forme $\frac{k}{n}$ avec $k \in \{0;1;2;\ldots;n\}$ $m_{dr}(\alpha)$ est la ***masse (en pourcentage)***
‖ ***possédée par les α derniers individus***.
4) ‖ La représentation graphique C_{dr} de m_{dr} est la ligne brisée passant pour $k = 0,1,\ldots,p$
‖ par les points $M'_k(1-\alpha_k, 1-\beta_k)$; c'est la symétrique de la courbe de Lorentz (C_{pr})
‖$(1-\alpha_k)$ et $(1-\beta_k)$ pour $k = 0,1,\ldots,p-1$
‖ sont respectivement les effectifs et les masses en cumulés décroissants (et en pourcentage).

Preuve :

1) La convexité de m_{pr} est prouvée en annexe 2 où on montre aussi que sur $[\alpha_i;\alpha_{i+1}]$ $m_{pr}(\alpha) = a_i\alpha + b_i$ avec $a_i = \frac{x_{i+1}}{\bar{x}}$ et $b_i = \beta_i - \frac{x_{i+1}}{\bar{x}}\alpha_i$ et donc $m_{pr}(\alpha) = \frac{x_1}{\bar{x}}\alpha$ pour $\alpha \in [0;\alpha_1]$ puisque $\alpha_0 = \beta_0 = 0$.
Soit $\alpha = \frac{k}{n} \in [\alpha_i;\alpha_{i+1}]$: on a alors $n_1 + n_2 + \ldots + n_i \leq k \leq n_1 + n_2 + \ldots + n_{i+1}$ et donc les k premiers individus sont constitués des $n_1 + n_2 + \ldots + n_i$ premiers individus, lesquels possédent (en pourcentage) β_i (voir définition de β_i au 3.1.1) et des $k - n_1 + n_2 + \ldots + n_i$ individus suivants qui ont comme valeur du caractère x_{i+1} et donc possédent $\frac{(k-(n_1+n_2+\ldots+n_i))x_{i+1}}{m} = \frac{n(\alpha-\alpha_i)}{m}x_{i+1} = (\alpha-\alpha_i)\frac{x_{i+1}}{\bar{x}}$.
Finalement la masse possédée par les $\frac{k}{n}$ premiers individus est $\beta_i + (\alpha-\alpha_i)\frac{x_{i+1}}{\bar{x}} = m_{pr}(\alpha)$.

2) Soit C_{pr} la courbe de Lorentz, c'est à dire la représentation graphique de m_{pr}, et C_{dr} sa symétrique par rapport au milieu de $[OA]$, milieu dont les coordonnées sont $(\frac{1}{2},\frac{1}{2})$. Pour tout $\alpha \in [0;1]$ le point $M(1-\alpha, m_p(1-\alpha))$ est sur C_{pr}, or son symétrique, qui a pour coordonnées $(\alpha, 1-m_{pr}(1-\alpha))$ est sur C_{dr} donc $1 - m_{pr}(1-\alpha) = m_{dr}(\alpha)$.

3) La concavité de m_{dr} est évidemment une conséquence de la convexité de m_{pr} et de la symétrie des deux représentations graphiques. Précisons cependant les calculs.
Soient u, α_1, α_2 entre 0 et 1 :

$$\begin{aligned} m_{pr}(1-(u\alpha_1+(1-u)\alpha_2) &= m_{pr}(u(1-\alpha_1)+(1-u)(1-\alpha_2)) \\ &\leq um_{pr}(1-\alpha_1)+(1-u)m_{pr}(1-\alpha_2) \end{aligned}$$

d'où

$$\begin{aligned} m_{dr}(u\alpha_1+(1-u)\alpha_2) &= 1-m_{pr}(1-(u\alpha_1+(1-u)\alpha_2) \\ &\geq 1-um_{pr}(1-\alpha_1)-(1-u)m_{pr}(1-\alpha_2) \\ &\geq u(1-m_{pr}(1-\alpha_1))\ +(1-u)(1-m_{pr}(1-\alpha_2)) \\ &\geq um_{dr}(\alpha_1)+(1-u)m_{dr}(\alpha_2) \end{aligned}$$

et donc m_{dr} est bien concave.
Si $\alpha \in [0;1-\alpha_{p-1}]$ alors $1-\alpha \in [\alpha_{p-1};1]$ d'où
$m_{dr}(\alpha) = 1-m_{pr}(1-\alpha) = 1-(a_{p-1}(1-\alpha)+b_{p-1}) = \alpha a_{p-1}+1-a_{p-1}-b_{p-1}$.
Compte tenu de la valeurs de a_i (voir le 1 ci-dessus) on obtient $a_{p-1} = \frac{x_p}{\bar{x}}$; quant à $1-a_{p-1}-b_{p-1}$ il est nul d'aprés l'annexe 2 et donc on a bien $m_{dr}(\alpha) = \alpha\frac{x_p}{\bar{x}}$.
Enfin, la relation $m_{dr}(\alpha) = 1-m_{pr}(1-\alpha)$ prouve que $m_{dr}(\alpha)$ est la masse en pourcentage possédée par les α derniers individus.

4) Résulte du fait que la courbe de Lorentz (C_{pr}) passe par les points $M_k(\alpha_k,\beta_k)$ dont les symétriques par rapport à $(\frac{1}{2},\frac{1}{2})$ sont les points $M'_k(1-\alpha_k,1-\beta_k)$.
Illustration :

4.3 Propriété

1) ‖ $\forall\alpha \in [0;1]\ m_{pr}(\alpha) \leq \alpha \leq m_{dr}(\alpha)$
2) ‖ Il existe $\alpha \in]0;1[$ tel que $m_{pr}(\alpha) = \alpha \Leftrightarrow$ il y a répartition égalitaire.
   ‖ Cela signifie que dès que la courbe de lorentz touche le segment $[OA]$ en un point
   ‖ autre que O et A alors, elle est entièrement confondue avec $[OA]$ et donc pour tout
   ‖ α dans $[0;1]$ on a $m_{pr}(\alpha) = \alpha$.
3) ‖ Il existe $\alpha \in]0;1[$ tel que $m_{dr}(\alpha) = \alpha \Leftrightarrow$ il y a répartition égalitaire.

Preuve :

1) $m_{pr}(\alpha) \leq \alpha$ résulte du fait que la courbe de Lorentz est entiérement située sous le segment $[OA]$
L'autre inégalité résultant du fait que la symétrique de la courbe de Lorentz est au dessus du segment $[OA]$, ou on peut dire que $m_{pr}(1-\alpha) \leq 1-\alpha$ et $m_{dr}(\alpha) = 1 - m_{pr}(1-\alpha)$ donne $\alpha \leq m_{dr}(\alpha)$.

2) Si $p > 1$ on a vu au 3.1.3 que la courbe de Lorentz C_{pr} est strictement en dessous de $[OA]$, sauf pour les points O et A et donc en fait dans ce cas, on a $\forall \alpha \in]0;1[\ m_{pr}(\alpha) < \alpha$.
Par contre si $p = 1$ il y a répartition égalitaire et $C_{pr} = [OA]$ c'est à dire $\forall \alpha \in [0;1]$ on a $m_{pr}(\alpha) = \alpha$.
Donc s'il existe $\alpha \in]0;1[$ tel que $m_{pr}(\alpha) = \alpha$, p ne peut être supérieur à 1, donc il est égal à 1 et la répartition est égalitaire (voir 2.2.5). La réciproque est évidente.
En fait on peut démontrer directement :
si f est une fonction convexe sur $[0;1]$ avec $f(0) = 0$ et $f(1) = 1$
alors a) $\forall x \in [0;1]$ on a $f(x) \leq x$
b) $\forall x \in [0;1]$ $f(x) = x \Leftrightarrow$ il existe $x \in]0;1[$ tel que $f(x) = x$.

3) Résulte de $m_{dr}(\alpha) = \alpha \Leftrightarrow m_{pr}(1-\alpha) = 1-\alpha$ et comme $1-\alpha \in]0;1[$ on applique 2.

Remarque 1 :
Soit α fixé dans $]0;1[$; $m_{pr}(\alpha)$ ne peut prendre la valeur α que s'il y a répartition égalitaire : on a donc ici une caractérisation simple de la répartition égalitaire. ***Pourquoi ne pas prendre comme indicateur de concentration*** $m_{pr}(\alpha)$? Ce point sera éxaminé plus loin.

Remarque 2 :
Une démonstration géométrique du 2 ci-dessus est donnée dans la remarque 2 de la propriété suivante.

4.4 Propriété

1) ‖ Si E, F, G sont 3 points d'une courbe de Lorentz dont les abcisses vérifient
‖ $0 \leq x_E < x_F < x_G \leq 1$ alors on a $pente[EF] \leq pente[EG] \leq pente[FG]$
‖ L'inégalité $pente[EF] \leq pente[EG]$ généralise le 4 du 3.1.3
‖ L'inégalité $pente[EF] \leq pente[FG]$ généralise le 6 du 3.1.3
2) ‖ Si U, V sont 2 points d'une courbe de Lorentz dont les abscisses vérifient
‖ $0 < x_U < x_V < 1$ alors on a $pente[OU] \leq pente[UV] \leq pente[VA]$
3) ‖ Soient U, V 2 points tels que
‖ $0 < x_U < x_V < 1$, $0 < y_U$, donc cf le 2), $pente[OU] \leq pente[UV] \leq pente[VA]$
‖ On note I le point d'intersection de $[OB]$ et (UV)
‖ On note J le point d'intersection de (OU) et (AV)
‖ On note K le point d'intersection de (UV) et $[AB]$

‖ On a alors le résultat suivant : toute courbe C_{pr} de Lorentz passant par les points U et V vérifie :
‖ entre les abcisses 0 et x_U C_{pr} est à l'intérieur du triangle OIU
‖ entre les abcisses x_U et x_V C_{pr} est à l'intérieur du triangle UJV
‖ entre les abcisses x_V et 1 C_{pr} est à l'intérieur du triangle VKA
‖ Le résultat s'adapte si $U = V$, auquel cas on supprime l'hypothése sur la pente $[UV]$,
‖ on supprime la définition de J, I est le point d'intersection de $[OB]$ et (AU), K est le point
‖ d'intersection de $[AB]$ et (OU) et alors on a :
‖ entre les abcisses 0 et x_U C_{pr} est à l'intérieur du triangle OIU
‖ entre les abcisses x_U et 1 C_{pr} est à l'intérieur du triangle UKA
‖ (voir figure correspondante à la remarque 2)

Preuve :

1) Soit G_1 le groupe des $x_G - x_E$ individus situés ‹‹après›› les x_E premiers individus (c'est à dire G_1 est le groupe des x_G premiers individus privé du groupe des x_E premiers individus), soit G_2 le groupe des $x_F - x_E$ individus situés après les x_E premiers individus et soit G_3 le groupe des $x_G - x_F$ individus situés après les x_F premiers individus.
On a $G_1 = G_2 \cup G_3$ et les individus de G_2 ont tous une valeur du caractère inférieure ou égale à toutes les valeurs du caractère prises par les individus de G_3 donc

$$\bar{x}_{G_2} \leq \bar{x}_{G_1} \leq \bar{x}_{G_3}$$
$$mse(G_2) \leq mse(G_1) \leq mse(G_3)$$
$$\frac{y_F - y_E}{x_F - x_E} \leq \frac{y_G - y_E}{x_G - x_E} \leq \frac{y_G - y_F}{x_G - x_F}$$
$$pente[EF] \leq pente[EG] \leq pente[FG]$$

2) On applique 1) avec $E = O$, $F = U$, $G = V$ d'où pente$[OU] \leq$ pente$[UV]$, puis on réapplique 1) avec $E = U$, $F = V$, $G = A$ d'où pente$[UV] \leq$ pente$[VA]$.

3) ***Montrons que*** I ***est sur*** $[OB[$, K ***sur*** $[AB[$ ***et*** U, V, J ***dans le triangle*** OBA
Si I est à gauche de O alors
$pente[UV] = pente[IU] = \frac{y_U - y_I}{x_U - x_I} = \frac{y_U}{x_U - x_I} < \frac{y_U}{x_U} = pente[OU]$
d'où $pente[UV] < pente[OU]$ ce qui est exclu.
Si I est à droite de B (ou $= B$) alors $pente[IU] = pente[UV]$ serait négative ce qui est exclu car $pente[OU] = \frac{y_U}{x_U} > 0$ et $pente[OU] \leq pente[UV]$.
Si K est en dessous de B (ou $= B$) alors $pente[UK] = pente[UV]$ serait négative (car y_K serait ≤ 0 alors que $y_U > 0$ et donc $y_K - y_U$ serait < 0)ce qui est exclu (voir ci-dessus).

Si K est en dessus de A alors $pente[VK] = pente[UV]$ serait supérieure à $pente[VA]$ ce qui est exclu.

U est dans le triangle OBA car son abcisse étant inférieure à 1 et son ordonnée supérieure à 0 il est forcément sur le segment $]IK[$ lequel est dans OBA ; quant à V, compte tenu de son abcisse qui est entre x_U et 1 il est à l'intérieur du segment $]UK[$ donc dans OBA et enfin J est dans OBA car les segments $[OU]$ et $[AV]$ sont dans OBA.

Justifions qu'il existe au moins une courbe de Lorentz C_{pr} passant par les points U et V

Si $pente[OU] = pente[UV] = pente[VA]$ $C_{pr} = [OA]$ convient.

Si $pente[OU] = pente[UV] \leq pente[VA]$ la ligne polygonale reliant O et A avec pour seul point anguleux le point V est une courbe de Lorentz qui convient (voir 3.3).

Si $pente[OU] < pente[UV] = pente[VA]$ la ligne polygonale reliant O et A avec pour seul point anguleux le point U est une courbe de Lorentz qui convient (voir 3.3).

Si $pente[OU] < pente[UV] < pente[VA]$ la ligne polygonale reliant O et A ayant pour seuls points anguleux les points U et V est une courbe de Lorentz qui convient (voir 3.3).

Montrons maintenant le résultat annoncé pour toute courbe de Lorentz C_{pr} passant par U et V.

Par exemple montrons que si $M(\alpha, \beta)$ est un point de C_{pr} avec $x_U < \alpha < x_V$ alors le point M est effectivement à l'intérieur du triangle UJV :

a) on applique 1 pour les 3 points U, M, V ce qui donne $pente[UM] \leq pente[UV] \leq pente[MV]$

b) on applique 1 pour les 3 points O, U, M ce qui donne $pente[OU] \leq pente[OM] \leq pente[UM]$

c) on applique 1 pour les 3 points M, V, A ce qui donne $pente[MV] \leq pente[MA] \leq pente[VA]$

Le a) et b) permettent d'écrire :

$$pente[OU] \leq pente[UM] \leq pente[UV]$$

$$\text{soit } pente[UJ] \leq pente[UM] \leq pente[UV]$$

et donc $[UM]$ est donc ‹‹entre›› les demi-droites $[UJ)$ et $[UV)$.

Le a) et b) permettent d'écrire :

$$pente[UV] \leq pente[MV] \leq pente[VA]$$

$$\text{soit } pente[VU] \leq pente[VM] \leq pente[VJ]$$

et donc $[VM]$ est donc ‹‹entre›› les demi-droites $[VJ)$ et $[VU)$ et M est bien dans le triangle OBA.

<u>*Remarque 1*</u> :

Si $pente[OU] = pente[UV]$, c'est-à-dire si O, U, V sont alignés alors $I = O$, $J = V$ et les triangles OIU et UJV sont aplatis. Dans ce cas le fait que C_{pr} soit à l'intérieur du triangle OIU (pour $0 \leq x \leq x_U$) et que C_{pr} soit à l'intérieur du triangle UJV (pour $0 \leq x_U \leq x_V$) signifie que la courbe de Lorentz C_{pr} est égale au segment $[OV]$ pour $0 \leq x \leq x_V$.

<u>*Remarque 2*</u> :

on verra à l'annexe 8 une application de 4.4 donnant un encadrement de g en fonction de $m_{pr}(0,5)$ et $m_{pr}(0,1)$ en prenant pour points U et V les points de la courbe de Lorentz d'abscisses respectives 0,5 et 0,9.

<u>*Remarque 3*</u> :

Dans le cas où $U = V$, toute courbe C_{pr} de Lorentz (qui passe par U) est donc dans la zone coloriée ci-dessous (cad à l'intérieur des triangles OIU et UKA) :

Note : $O(0,0), I(0,4,0), U \simeq (0,64,0,36), K(1,0,6), B(1,1$

On observe alors que plus l'ordonnée de U augmente, plus I et K se rapprochent respectivement de O et A et plus C_{pr} sera proche de $[OA]$; et dés que U sera sur $[OA]$ on aura $I = O$ et $K = A$ et donc $C_{pr} = [OA]$: on retrouve ici la propriété 4.3 précédente, à savoir que dés qu'un point de la courbe de Lorentz est sur $]OA[$, alors la courbe de Lorentz est entièrement confondue avec $[OA]$.

4.5 Propriété

1) $\| \; g = 2\int_0^1 (\alpha - m_{pr}(\alpha))d\alpha = 1 - 2\int_0^1 m_{pr}(\alpha)d\alpha$
   $\| \; g = 2\int_0^1 (m_{dr}(\alpha) - \alpha)d\alpha = -1 + 2\int_0^1 m_{dr}(\alpha)d\alpha$
2) $\| \; \frac{1}{2}g$ est la valeur moyenne des écarts effectifs moins masse possédée pour tous les groupes
   $\|$ constitués des premiers individus.
   $\| \; \frac{1}{2}g$ est la valeur moyenne des écarts masse possédée moins effectifs pour tous les groupes
   $\|$ constitués des derniers individus.
3) $\| \; 2g\frac{m}{n}$ est la valeur moyenne de la variable aléatoire égale à la valeur absolue de la différence
   $\|$ des valeurs du caractère de deux individus choisis au hasard, parmi les n de la population.

Preuve :

1) g est par définition deux fois l'aire de la région comprise entre $[OA]$ et la courbe C_{pr} de Lorentz ce qui donne la première égalité. Mais par raison de symétrie cette aire est aussi celle comprise entre C_{dr} et $[OA]$ ce qui donne la troisième égalité.

2) Résulte du fait que les α premiers individus possédent $m_{pr}(\alpha)$ (voir 4.2) et de :
si f est une fonction continue sur $[a;b]$
alors $\frac{1}{b-a}\int_a^b f(x)dx = \lim_{k\to+\infty} \frac{f(a) + f(a + \frac{b-a}{k}) + f(a + 2\frac{b-a}{k}) + \ldots\ldots f(a + (k-1)\frac{b-a}{k})}{k}$
et donc $\frac{1}{b-a}\int_a^b f(x)dx$ est appelée valeur moyenne de f sur $[a;b]$.

3) Voir la preuve en annexe 5. Ce résultat a été mis ici pour son aspect valeur moyenne, tout comme le résultat 2) précédent, même s'il est de nature différente et qu'il ne sera pas exploité, lui, par la suite.

Remarque :

Cette interprétation intégrale de g permet de retrouver des résultats déjà obtenus :
1) $g \in [0;1]$ puisque $0 \leq \alpha - m_{pr}(\alpha) \leq \alpha$ et on intégre entre 0 et 1.
2) $g = 1$ est impossible car $\forall \alpha \in [0;1] \; m_{pr}(\alpha) \geq 0$ donc $\int_0^1 m_{pr}(\alpha)d\alpha \geq 0$; mais la fonction m_{pr} est continue avec

$m_{pr}(1) = 1 > 0$ donc $\int_0^1 m_{pr}(\alpha)d\alpha > 0$ et $g < 1$.

3) $g = 0 \Leftrightarrow$ la répartition est égalitaire : en effet $g = 0$ entraîne $\int_0^1(\alpha - m_{pr}(\alpha))d\alpha = 0$ or l'expression $\alpha - m_{pr}(\alpha)$ est positive ou nulle sur $[0;1]$ et est continue sur cet intervalle, donc son intégrale ne pourra être nulle que si et seulement si cette expression est toujours nulle, c'est à dire si $\forall\alpha \in [0;1]$ on $\alpha = m_{pr}(\alpha)$. Donc $g = 0$ entraîne que la répartition est égalitaire (voir 4.3). La réciproque est immédiate : si la répartition est égalitaire alors $p = 1$, C_{pr} est confondue avec $[OA]$ et $\forall\alpha \in [0;1]$ on $\alpha = m_{pr}(\alpha)$ donc $g = 0$.

4) On peut aussi retrouver la formule du 3.1.4 : on part de $g = 1 - 2\int_0^1 m_{pr}(\alpha)d\alpha$ et on calcule l'intégrale en la découpant en p intégrales sur $[\alpha_i, \alpha_{i+1}]$ pour i allant de 0 à $p-1$ et on utilise le fait que sur un tel intervalle $m_{pr}(\alpha) = a_i\alpha + b_i$ avec $a_i = \frac{x_{i+1}}{\bar{x}}$ et $b_i = \beta_i - \frac{x_{i+1}}{\bar{x}}\alpha_i$ (voir annexe 2).

4.6 Propriété

‖ $\frac{4}{3}(m_{dr}(0,5) - 0,5) = \frac{2}{3}(\frac{m_{dr}(0,5)}{0,5} - 1)$ est une valeur approchée du coefficient de Gini.

Preuve :

On applique la méthode de Simpson (référence [4] page 200) permettant de calculer une valeur approchée d'une intégrale :

$$\int_0^1 m_{dr}(\alpha)d\alpha \simeq \frac{1-0}{6}(m_{dr}(0) + m_{dr}(1) + 4m_{dr}(0,5)) \text{ et donc}$$
$$g \simeq -1 + \frac{2}{6}(1 + 4m_{dr}(\alpha)) = \frac{4}{3}m_{dr}(0,5) - \frac{2}{3}.$$

Remarque :

Evidemment la précision de cette approximation est variable : elle peut être correcte lorsque la courbe de Lorentz a l'allure d'une parabole ou d'une cubique, puisque la méthode de Simpson donne un résultat exact pour les polynômes de degré inférieur ou égal à 3 (c'est le cas de l'exemple 6.1 où $g \simeq 0,427$ et $\frac{4}{3}(m_{dr}(0,5) - 0,5) \simeq 0,39)$, par contre lorsque C_{pr} n'a ni l'allure d'une parabole, ni l'allure d'une cubique, et c'est notamment le cas lorsque C_{pr} est proche de $[OB] \cup [BA]$ (cad lorsque $g \simeq 1$, cf 3.1.5), l'approximation est alors de moins bonne qualité (voir l'exemple 6.4 où $g \simeq 0,8476$ et $\frac{4}{3}(m_{dr}(0,5) - 0,5) \simeq 0,64$).

Mais le but n'est pas ici de chercher une bonne valeur approchée de g : ***l'intérêt de cette approximation est en fait de montrer que bien souvent le coefficient de Gini est pratiquement égal, à une transformation affine prés, à la masse possédée par les 50% derniers individus (ou à la masse possédée par les 50% premiers individus puisque la somme de ces deux masses est 1)***.

5 Un indicateur de concentration e équivalent au coefficient de Gini mais immédiat à calculer et avec une interprétation économique précise.

Le coefficient de Gini mesure en fait la proximité ou non de la courbe de Lorentz C_{pr} au segment $[OA]$: si $g = 0$ cette courbe C_{pr} est confondue avec $[OA]$, et plus g est proche de 1 ($g = 1$ est impossible) plus C_{pr} s'éloigne de $[OA]$. Or le raisonnement fait au cours de la démonstration de la propriété 3.2.1 montre que la quantité $\frac{\sqrt{2}}{2}(\alpha_k - \beta_k)$ est la distance (euclidienne) entre le point $M_k(\alpha_k, \beta_k)$ et le segment $[OA]$, la quantité $(\alpha_k - \beta_k)$ étant la distance, parallélement à (OB), entre le point M_k et le segment $[OA]$.

La valeur maximum des $\alpha_k - \beta_k$ (écart maximum entre l'effectif et la masse possédée par le groupe des α_k premiers individus) ***est donc, elle aussi, une mesure de la proximité de la courbe de Lorentz au segment*** $[OA]$.

Intéressons nous donc à cette valeur maximum qui, on va le voir, a pratiquement les mêmes propriétés que g tout en étant immédiate à calculer (il suffira de créer, aprés les 2 colonnes des effectifs et masses en cumulés croissants, la colonne des différences : on verra au chapitre suivant des exemples détaillés.)

5.1 Définition

‖ On note $e = \max_{k \in \{0;1;....p\}} (\alpha_k - \beta_k)$

‖ Bien entendu $e = 0$ si $p = 1$ car $\alpha_0 - \beta_0 = \alpha_1 - \beta_1 = 0$.

‖ Si $p \geq 2$ alors $e = \max_{k \in \{1;....p-1\}} (\alpha_k - \beta_k) > 0$

‖ car $\alpha_0 - \beta_0 = \alpha_p - \beta_p = 0$ et $\alpha_k - \beta_k > 0$ pour $k = 1, 2, \ldots, p-1$

Précisons les liens entre g et e ;

5.2 Propriété

1) ‖ $e = \max_{\alpha \in [0;1]} (\alpha - m_{pr}(\alpha))$: c'est à dire e est la valeur maximum des écarts effectifs moins
   ‖ masse possédée pour tous les groupes constitués des α premiers
   ‖ individus alors que $\frac{1}{2}g$ est la valeur moyenne de ces écarts.
2) ‖ $e = \max_{\alpha \in [0;1]} (m_{dr}(\alpha) - \alpha)$
3) ‖ Si $p = 1$ alors $e = g = 0$ et si $p = 2$ alors $e = g = \alpha_1 - \beta_1$
4) ‖ Tout comme g, $e \in [0;1[$ et sur l'ensemble de toutes les séries e décrit tout $[0;1[$
5) ‖ Tout comme g , e mesure la proximité de la courbe C_{pr} de Lorentz et du segment $[OA]$:
   ‖ la courbe C_{pr} est située entre la droite (OA) d'équation $y = x$ et la droite d'équation $y = x - e$
   ‖ ces 2 droites étant paralléles et distantes de $\frac{\sqrt{2}}{2}e$.
6) ‖ Si $p \geq 2$ et si i est tel que $e = \alpha_i - \beta_i$ (i est forcément différent de 0 et 1) alors
   ‖ $g - e =$ deux fois l'aire de la région située entre la courbe de Lorentz et la ligne
   ‖ polygonale OM_iA.
7) ‖ $\frac{g}{2} \leq e \leq g$

8) ‖ Tout comme g on a $e \leq \dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}} < 1$

‖ $\dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ est un majorant de e pour toutes les séries à x_1 et x_p fixés

‖ Ce majorant est atteint si et seulement si $p = 2$ et $\dfrac{n_1}{n} = \dfrac{\sqrt{x_2}}{\sqrt{x_2} + \sqrt{x_1}}$

9) ‖ $e = 0 \Leftrightarrow g = 0$ et e ‹‹proche›› de $1 \Leftrightarrow g$ ‹‹proche›› de 1

10) ‖ Comme pour g, $e \simeq 0$ ne signifie pas qu'il y a répartition presque égalitaire (voir 2.2.8)

Preuve :

1) Si $p = 1$ c'est évidemment vrai puisque $e = 0$ et $\forall \alpha \in [0;1]\ \alpha - m_{pr}(\alpha) = 0$.
Supposons $p \geq 2$ et rappelons que $\beta_k = m_{pr}(\alpha_k)$ pour tout entier k de 0 à p.
S'il existe $u \in [0;1]$ avec $u - m_{pr}(u) > e$ alors forcément u est différent de tous les α_k pour $k = 0, 1, \ldots, p$ et donc il existe $i \in \{0; 1; \ldots; p-1\}$ tel que $u \in]\alpha_i; \alpha_{i+1}[$.
La fonction h définie par $h(\alpha) = \alpha - m_{pr}(\alpha)$ étant affine sur $[\alpha_i; \alpha_{i+1}]$ comme différence de 2 fonctions affines (voir 4.2) il y a alors 3 possibilités sur $[\alpha_i; \alpha_{i+1}]$:
soit h est constante et $h(\alpha_i) = h(u)$, d'où $\alpha_i - \beta_i > e$ ce qui est impossible.
soit h est strictement croissante et $h(u) < h(\alpha_{i+1})$ d'où $e < \alpha_{i+1} - \beta_{i+1}$ encore impossible.
soit h est strictement décroissante et $h(u) < h(\alpha_i)$ d'où $e < \alpha_i - \beta_i$ encore impossible.
Finalement il ne peut exister u dans $[0;1]$ avec $u - m_{pr}(u) > e$ et donc $\forall u \in [0;1]$ on a $u - m_{pr}(u) \leq e$, ce qui prouve le résultat.

2) $\max\limits_{\alpha \in [0;1]} m_{dr}(\alpha) - \alpha = \max\limits_{\alpha \in [0;1]} m_{dr}(1-\alpha) - (1-\alpha) = \max\limits_{\alpha \in [0;1]} \alpha - m_{pr}(\alpha)$
(puisque $m_{dr}(1-\alpha) = 1 - m_{pr}(\alpha)$).

3) Si $p = 2$ alors $g = \alpha_1 - \beta_1$ d'aprés 3.2.1 et $e = \max\limits_{k \in \{1\}} \alpha_k - \beta_k = \alpha_1 - \beta_1$

4) Si $p = 1$ $e = 0 \in [0;1[$
Si $p \geq 2$ $e = \alpha_i - \beta_i$ avec $i \in \{1; 2; \ldots; p-1\}$, et donc $\alpha_i < 1$ (seul $\alpha_p = 1$) et comme $\beta_i > 0$ on a bien $e \in [0;1[$.
Notons que ce résultat est encore vrai si $x_1 = 0$ et $p \geq 2$, auquel cas on a $\beta_1 = 0$, mais même si $e = \alpha_1 - \beta_1$ on aura encore $e < 1$ car $\alpha_1 < 1$.
Enfin e décrit tout $[0;1[$ car si $p = 2$ alors $e = \alpha_1 - \beta_1$, quantité qui décrit tout $]0;1[$ (voir preuve du 4 de 3.2.3).

5) La définition de e et le fait que tout point $M_k(\alpha_k, \beta_k)$ est sur la droite d'équation $y = x + \beta_k - \alpha_k$ entraîne que la courbe de Lorentz est effectivement comprise entre la droite (OA) et la droite $y = x - e$, lesquelles sont distantes de $\dfrac{\sqrt{2}}{2}e$ puisque la distance du point $M_k(\alpha_k, \beta_k)$ à la droite (OA) est $\dfrac{\sqrt{2}}{2}(\alpha_k - \beta_k)$ (voir la démonstration de 3.2.1).
Il est à noter que la courbe de Lorentz a un point commun avec la droite d'équation $y = x - e$: c'est le point $M_i(\alpha_i, \beta_i)$ tel que $e = \alpha_i - \beta_i$

Illustration :

Il est alors clair que plus e diminue, plus la courbe de Lorentz se rapproche de $[OA]$, et plus e augmente plus la courbe de Lorentz s'éloigne de $[OA]$: e est bien une mesure de la proximité de la courbe de Lorentz et du segment $[OA]$.

6) Soit i tel que $e = \alpha_i - \beta_i$: le triangle OM_iA est situé entre $[OA]$ et la courbe C_{pr} (voir figure ci-dessus) et donc le résultat annoncé est alors une conséquence immédiate de $\alpha_i - \beta_i = 2 \times$ aire du triangle OM_iA (voir 3.2.1) et de g =2×aire de la région située entre $[OA]$ et la courbe C_{pr}.

7) On a $g = 2\int_0^1 (\alpha - m_{pr}(\alpha))d\alpha$ (voir 4.4) et d'aprés le 1 ci-dessus :

$$g \leq 2\int_0^1 e d\alpha = 2e.$$

Le 6 ci-dessus prouve que $g - e \geq 0,$ soit $g \geq e$.

8) Le 7 ci-dessus et 3.2.3 permettent d'écrire $e \leq g \leq \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ d'où la majoration annoncée.

L'égalité $e = \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ entraîne $g = \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ ce qui exige, d'aprés 3.2.3, que $p = 2$ et $\frac{n_1}{n} = \frac{\sqrt{x_2}}{\sqrt{x_2} + \sqrt{x_1}}$.

Réciproquement si $p = 2$ et $\frac{n_1}{n} = \frac{\sqrt{x_2}}{\sqrt{x_2} + \sqrt{x_1}}$ alors $g = \frac{\sqrt{x_2} - \sqrt{x_1}}{\sqrt{x_2} + \sqrt{x_1}}$ (voir 3.2.3) mais on est dans le cas où $e = g$ et donc $e = \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$.

Une autre preuve peut être faite sans utiliser la majoration de g : on utilise 7.6 qui permet d'écrire que $m_{dr}(\alpha) - \alpha \leq f(\alpha)$ avec $f(\alpha) = \frac{\alpha x_p}{\alpha x_p + (1-\alpha)x_1} - \alpha$, puis par étude du sens de variation de f on montre que f a une valeur maximum $f(\frac{\sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}) = \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$, ce qui donne la majoration de e. On aura $e = \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ si et seulement si il existe $\alpha \in [0;1]$ tel que $m_{dr}(\alpha) = \frac{\alpha x_p}{\alpha x_p + (1-\alpha)x_1}$ et $f(\alpha) = \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$, c'est à dire si et seulement si $p = 2$, $\alpha = \frac{n_2}{n}$ (voir 4 du 7.6) et $\alpha = \frac{\sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ ce qui redonne $\frac{n_1}{n} = \frac{\sqrt{x_2}}{\sqrt{x_p} + \sqrt{x_1}}$ (puisque $\frac{n_1}{n} + \frac{n_2}{n} = 1$).

9) Le résultat 7 entraîne la première équivalence et le fait que si e est proche de 1 alors il en est de même pour g.
Si maintenant c'est g qui est proche de 1 alors C_{pr} est proche de $[OB] \cup [BA]$ et il existe i tel que le point $M_i(\alpha_i, \beta_i)$ soit proche de B, donc il existe i tel que $\alpha_i \simeq 1$ et $\beta_i \simeq 0$ soit $\alpha_i - \beta_i \simeq 1$ et comme e est la valeur maximum des $\alpha_k - \beta_k$ et qu'il est inférieur à 1, forcément $e \simeq 1$.

10) Si $p = 2$ alors $e = g$ et on a vu que dans cas $g \simeq 0$ n'entraîne pas qu'il y a répartition égalitaire (voir 3.2.2)..

Remarque :
Sur l'ensemble des quelques exemples traités au chapitre suivant, l'ordre de grandeur de $g - e$ est 0,1.

5.3 Commentaires sur ce nouveau coefficient e

Ce coefficient** e **se comporte exactement comme le coefficient de Gini puisque les résultats 4 et 9 de la propriété précédente permettent de dire :

$\| \; e \in [0;1[$
$\| \; e = 0$ signifie qu'il y a répartition égalitaire
$\|$ (puisque $e = 0 \Leftrightarrow g = 0$)
$\| \; e > 0$ signifie qu'il y a concentration
$\|$ (puisqu'il n'y a pas répartition égalitaire ; voir 2.2.2)
$\| \; e \simeq 1$ signifie que beaucoup (les premiers) ont peu.
$\|$ (puisque $e \simeq 1 \Leftrightarrow g \simeq 1$)

Remarquons que ce dernier résultat peut se démontrer directement : si $e \simeq 1$, c'est qu'il existe i tel que $\alpha_i - \beta_i \simeq 1$, et comme $0 \leq \beta_i \leq \alpha_i \leq 1$ on a $\alpha_i \simeq 1$ et $\beta_i \simeq 0$, c'est à dire les α_i premiers individus, donc beaucoup, ont peu (β_i).

***Mais ce coefficient** e **présente deux avantages par rapport au coefficient de Gini** :*

1) ***Sa simplicité de calcul** :*
Il suffit de faire les différences $\alpha_k - \beta_k$! En fait il n'est même pas nécéssaire de calculer toutes ces différences puisque la suite $\alpha_k - \beta_k$ est d'abord croissante puis décroissante (voir annexe 1) et donc, on commence par calculer $\alpha_1 - \beta_1$, puis $\alpha_2 - \beta_2$ et on s'arrête dès que l'on constate la décroissance : le coefficient e est alors la différence précédente. On verra au chapitre suivant de nombreux exemples traités complétement.

2) ***Il a une interprétation économique trés précise** :*
e est la valeur maximum des écarts effectif moins masse possédée pour tous les groupes constitués de premiers individus, laquelle est égale à la valeur maximum des écarts masse possédée moins effectif pour tous les groupes constitués de derniers individus.

En outre le calcul permet de mettre en évidence le(s) groupe(s) réalisant cet écart maximum.

Par exemple si $e = 0,6$ c'est qu'il existe i tel que $\alpha_i - \beta_i = 0,6$ et donc le groupe des α_i premiers individus réalise cet écart maximum : il a un effectif supérieur de 0,6 à sa masse alors que le groupe des $1 - \alpha_i$ derniers individus a une masse supérieure de 0,6 à son effectif.

Par contre le fait de trouver $g = 0,7$ (0,7 car en moyenne, du moins sur les exemples du 6, $g - e \simeq 0,1$) ne permet pas d'en déduire quelque chose d'aussi précis : on peut simplement dire que 0,35 est la valeur moyenne des écarts effectif moins masse possédée pour tous les groupes constitués de premiers individus (voir 4.3).

***Cependant**, **ce coefficient a tout comme** g **deux inconvénients** :* $e \simeq 0$ peut cacher une répartition qui n'est pas presque égalitaire, c'est à dire il peut tout de même exister un groupe ayant beaucoup plus en masse qu'en effectif (voir 3.2.2), ***et surtout à un même** e **(ou à un même** g**) peut correspondre des courbes de Lorentz différentes donc des situations différentes en terme de concentration**.*

Illustration :

Ces deux courbes de Lorentz vérifient $e = g$ car $p = 2$ et la droite (M_1, N_1) est paralléle à (OA).

Examinons la courbe de Lorentz passant par le point $M_1(0,7\ ,0,1)$: cela signifie que les 30% derniers individus de cette série possèdent 90% de la masse totale et comme ils ont tous la même valeur du caractère (x_2) les 10% derniers individus possèdent 30% de la masse totale.

L'autre courbe de Lorentz passe par le point $N_1(0,9\ ,0,3)$: cette fois les 10% derniers individus de la série possèdent 70% de la masse totale.

Au niveau des 10% derniers individus il y a donc une différence très importante malgré le fait qu'il y ait le même e (ou le même g) : ceci prouve que la notion de concentration ne peut être quantifiée par un seul nombre.

Pourquoi ne pas dire ici que la deuxième série est nettement plus concentrée que la première au niveau des 10% derniers individus? Ce point sera approfondi à partir du chapitre 7, le chapitre 6 ci-après étant consacré à des exemples de calculs de g et e accompagnés d'analyses comparatives de ces deux coefficients.

6
Comparaison de g et e sur quelques exemples

Il s'agit, sur quelques exemples pris pour la plupart dans la littérature existante sur le sujet, de calculer les coefficients g et e et de procéder à quelques analyses au cours desquelles on montrera l'intérêt de non pas considérer les différences effectif moins masse possédée, comme le font g et e, mais plutôt les rapports masse sur effectif (*mse*) dont on a déjà vu certaines propriétés au chapitre 2.

La plupart de ces exemples seront repris au dernier chapitre pour être analysés à la l'aide d'une nouvelle méthode reposant justement sur la considération de ces rapports masse sur effectif (***la méthode*** *mse*).

6.1 Exemple 1

Source : référence [5] page 79.

On considére une population dont les $n = 80$ individus sont des terres agricoles, le caractère étudié etant la superficie en hectares, caractère qui prend $p = 5$ valeurs (les x_i) :

x_i	n_i	$n_i x_i$	α_i	β_i	$\alpha_i - \beta_i$
5	16	80	0,2	0,0375	0,1625
15	30	450	0,575	0,248	0,327
30	18	540	0,8	0,502	0,298
55	10	550	0,925	0,760	0,165
85	6	510	1	1	0
	$n = 80$	$m = 2130$			

Toutes les notations utilisées sont celles définies aux 2.1 et 3.1.1 ; en particulier n est l'effectif total, m est la masse totale, les α_i (resp β_i) sont les effectifs (resp les masses possédées) en pourcentage et cumulés croissants.

Bien entendu la dernière colonne est inutile si on se limite au calcul du coefficient de Gini : elle sert à calculer le coefficient $e = \max_{k \in \{1;\ldots;p-1\}} (\alpha_k - \beta_k)$.

Cette derniére colonne donne immédiatement $e = 0,327$.

Cela signifie que le groupe des 57,5% premiers individus a un effectif supérieur de 0,327 à la masse qu'il posséde et donc le groupe des 42,5% derniers individus posséde une masse supérieure à son effectif de 0,327, ce chiffre de 0,327 étant la valeur ***maximum*** des différences effectif moins masse pour tous les groupes constitués de premiers individus (ou la valeur maximum des différences masse moins effectif pour tous les groupes constitués de derniers individus).

On notera, comme cela a déjà été dit (voir 3.1.3) que la suite $\alpha_i - \beta_i$ est d'abord croissante puis décroissante.

Par contre le calcul de g à partir du tableau ci-dessus est beaucoup plus lourd :

D'aprés 3.1.4 (et en se rappelant que $\alpha_0 = \beta_0 = 0$) on a :

$$\begin{aligned} g &= 1 - \sum_{i=0}^{4} \frac{n_{i+1}}{n} (\beta_i + \beta_{i+1}) \\ &= 1 - \frac{16 \times 0,0375 + 30 \times (0,0375 + 0,248) + 18 \times (0,248 + 0,502) + \ldots}{80} \\ &= 1 - \frac{45,845}{80} \\ &\simeq 0,427 \end{aligned}$$

Notons que l'on a bien $g \leq \dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$ (voir 3.2.3) puisque $\dfrac{\sqrt{85} - \sqrt{5}}{\sqrt{85} + \sqrt{5}} \simeq 0,61$.

Quelle conclusion en tirer? L'auteur de la référence 4 d'où est tiré l'exemple conclu à une ‹‹concentration›› moyenne puisque 0,427 est à mi distance de 0 (valeur mimimun de g, cas de la répartition égalitaire) et de 1 (limite supérieure de g). Notons qu'il aurait été peut-être plus rigoureux de dire en dessous de la moyenne, puisque 0,427 est nettement

en dessous de 0,5.

Cependant cette conclusion me parait peu satisfaisante : une fois que l'on a dit ‹‹concentration›› moyenne (ou en dessous de la moyenne) que sait-on de plus sur la série? Rien. On n'a pas vraiment répondu à la question qui est de savoir s'il y a des groupes possédant beaucoup plus en masse qu'en effectif.

Par contre le coefficient $e = 0,327$ a une interprétation beaucoup plus concréte (voir plus haut) : mais je crois que l'on peut faire encore mieux.

En effet d'aprés 2.2.3, pour tous les groupes possibles d'individus de la série, la valeur la plus forte du rapport masse sur effectif (*mse*) est :

$$\frac{x_p}{\bar{x}} = \frac{x_5}{\bar{x}} = \frac{85}{\frac{2130}{80}} \simeq 3,19$$

et elle est réalisée par le groupe des n_5 derniers individus. C'est à dire le groupe des 6 derniers individus (ceux ayant 85 comme valeur du caractère) et dont l'effectif représente ici 7,5% du total posséde une masse presque 3,2 fois plus grande que son effectif et, aucun autre groupe n'a un rapport masse sur effectif supérieur à ce chiffre.

Ce renseignement me semble trés intéréssant : cependant on peut objecter que lorsque que l'on change de série, l'effectif $\frac{n_p}{n}$ (en pourcentage) des n_p derniers individus va aussi changer ce qui rendra les comparaisons difficiles.

Par contre on peut considérer le *mse* des 50% derniers individus, c'est-à-dire ici le *mse* des 40 derniers individus qui est égal à

$$\frac{\frac{(40-34)\times 15+540+550+510}{2130}}{0,5} = 1,58.$$

Cela signifie que le groupe des 50 derniers individus a une masse égale à 1,58 fois son effectif ; ce rapport pouvant varier entre 1 (les 50% derniers ont toujours au moins 50% de la masse, voir 4.3) et 2 ($\frac{100\%}{50\%}$ si les 50% derniers ont tout, possible que si $x_1 = 0$ et $p \geq 2$), pourquoi ne pas dire que la concentration, au niveau des 50% derniers, est moyenne? C'est tout de même plus rapide et plus parlant que de dire $g = 0,427$.

A titre d'illustration voici la courbe de Lorentz de cette série : c'est la ligne brisée reliant les points $M_i(\alpha_i, \beta_i)$ pour $i = 0,1,2,3,4,5$ avec $M_0(0,0) = O$ et $M_5(1,1) = A$:

6.2 Exemples 2 et 3

Il s'agit ici des séries décilées des Revenus et Patrimoines 1986 tirées de la référence [1] pages 107 et 109.

Chaque classe ayant un effectif de 10% de l'effectif total *n*, pour les deux séries on a $\alpha_i = \frac{i}{10}$ avec $i = 1,2,\ldots,10$ (voir 3.1.4) :

n° de la classe	part du revenu	β_i	$\alpha_i - \beta_i$	part du patrimoine	β_i	$\alpha_i - \beta_i$
1	2,2%	0,022	0,078	0,1%	0,001	0,099
2	3,8%	0,06	0,14	0,3%	0,004	0,196
3	4,9%	0,109	0,191	0,8%	0,012	0,288
4	6%	0,169	0,231	1,6%	0,028	0,372
5	7,2%	0,241	0,259	3,2%	0,06	0,44
6	8,8%	0,329	0,271	5,9%	0,119	0,481
7	10,6%	0,435	0,265	8,6%	0,205	0,495
8	12,6%	0,561	0,239	10,6%	0,312	0,488
9	16,1%	0,722	0,178	15,1%	0,462	0,438
10	27,8%	1	0	53,8%	1	0

On déduit immédiatement de ce tableau :

pour la série des Revenus : $e = 0,271$, c'est à dire les 40% derniers individus ont une masse supérieure de 0,271 à leur effectif (c'est l'écart maximum).

pour la série des Patrimoines : $e = 0,495$, c'est à dire les 30% derniers individus ont une masse supérieure de 0,495 à leur effectif (c'est l'écart maximum).

Il n'y a donc pas besoin de la courbe de Lorentz ni du coefficient de Gini pour constater, avec une quantification précise à l'appui, que l'aspect concentration est plus marqué chez les patrimoines que chez les revenus.

Evidemment le coefficient de Gini et la courbe de Lorentz vont traduire cet aspect mais cela sera moins immédiat à obtenir.

On a ici $g = \frac{9}{10} - \frac{1}{5}(\beta_1 + \beta_2 + \ldots + \beta_9)$, cela d'aprés le 3.1.4 ce qui donne :

pour les Revenus $g = 0,9 - \frac{2,648}{5} \simeq 0,37$

et pour les Patrimoines $g = 0,9 - \frac{1,203}{5} \simeq 0,66.$

Ces deux valeurs permettent de retrouver la conclusion précédente : l'aspect concentration est plus marqué chez les patrimoines que chez les revenus

Notons que les rapports $\frac{0,495}{0,271} \simeq 1,83$ et $\frac{0,66}{0,37} \simeq 1,78$ sont pratiquement égaux, ce qui confirme le fait qu'utiliser le coefficient *e* équivaut quantitativement à utiliser le coefficient *g* mais rappelons que *e* est plus facile à calculer que *g* et qu'il a une interprétation concréte (voir plus haut), que *g* n'a pas.

Par ailleurs, on peut remarquer que tout le monde s'accorde à considérer que la série des patrimoines est très concentrée alors que son coefficient de Gini est assez éloigné de 1 (limite supérieure de *g*) ce qui est tout de même un petit peu curieux. En réalité pour les séries décilées *g* ne peut excéder 0,9 (voir 3.2.3), mais cela ne change pas grand chose : 0,66 reste éloigné de 0,9.

Par contre si on calcule le *mse* des 50% derniers individus on s'aperçoit qu'il est égal à

$$\frac{5,9\% + 8,6\% + 10,6\% + 15,1\% + 53,8\%}{50\%} = 1,88$$

chiffre qui lui est très proche de la valeur maximum 2 ($\frac{100\%}{50\%}$, possible si $x_1 = 0$ et $p \geq 2$) et qui justifie davantage le qualificatif de série très concentrée.

La nouvelle méthode d'analyse d'une série, en terme de concentration, qui sera développée dans les chapitres suivants montrera le bien-fondé de ce point de vue.

A titre d'illustration voici la courbe de Lorentz de ces deux séries : c'est la ligne brisée reliant les points $M_i(\alpha_i, \beta_i)$ pour $i = 0, 1, 2, \ldots 10$ avec $M_0(0,0) = O$ et $M_{10}(1,1) = A$:

.

6.3 Exemple 4

Source : référence [3] page 786.

x_i	n_i	$n_i x_i$	α_i	β_i	$\alpha_i - \beta_i$
2,5	48	120	0,6	0,024	0,576
12,5	16	200	0,8	0,064	0,736
60	8	480	0,9	0,16	0,74
300	4	1200	0,95	0,4	0,55
750	4	3000	1	1	
	$n = 80$	$m = 5000$			

On a immédiatement: $e = 0,74$, c'est-à-dire la différence maximum masse moins effectif est de 0,74 réalisée par le groupe des 10% derniers individus. Notons que les 20% derniers individus possédent une masse supérieure à leur effectif de 0,736, pratiquement égal à 0,74.

Le coefficient de Gini $g = 1 - \frac{1}{80} \times \frac{1}{125}(48 \times 3 + 16(3 + 8) + 8(8 + 20) + 4(20 + 50) + 4(50 + 125))$

$= 1 - \frac{1524}{80 \times 125} = 0,8476.$

Cette valeur proche de 1 permet à l'auteur de la référence de dire qu'il y a forte ‹‹concentration››, ce que l'on pouvait très bien dire tout de suite à partir de la seule lecture du coefficient *e*.

Ces fortes valeurs de *g* et de *e* sont évidemment des traductions du fait que la courbe de Lorentz est très proche de $[OA] \cup [BA]$, cela parce que M_1 et M_2 ont des ordonnées très faibles (0,024 et 0,064) et M_3 et M_4 ont des abscisses supérieures ou égales à $0,9$.

On peut aussi dire que la courbe de Lorentz est très proche de $[OA] \cup [BA]$ car les 90% derniers ont peu (16%), c'est-à-dire les 10% derniers ont beaucoup (84%), ce qui correspond à la forte valeur de $e = 0,84 - 0,10$.

Cependant il me semble beaucoup plus intéréssant de considérer non pas la différence 0,84-0,10 mais le rapport $\frac{0,84}{0,10} = 8,4$, c'est à dire le *mse* des 10% derniers individus (voir 2.2.1).

Cette valeur de 8,4 pour le *mse* des 10% derniers individus est proche de la valeur maximum $\frac{100\%}{10\%} = 10$ (cas où les 10% derniers possédent tout, possible si $x_1 = 0$ et $p \geq 2$) : ***c'est cet aspect qui me semble justifier le mieux le qualificatif de forte concentration, du moins au niveau des 10% derniers.***

Cherchons le *mse* des 50% derniers individus : $\frac{\frac{5000 - 40 \times 2,5}{5000}}{0,5} = 1,96,$ valeur trés proche de la valeur maximum

2. Là encore on peut dire qu'il y a trés forte concentration (au niveau des 50% derniers individus) puisqu'ils possédent presque tout.

Terminons cet exemple par deux remarques :

1) Le plus fort *mse* pour cette série est (voir 2.2.3) le *mse* du groupe constitué des 4 derniers individus (5% de l'effectif) : $\frac{\frac{3000}{5000}}{\frac{5}{100}} = 12.$

2) On peut vérifier ici un des résultats du 2.2.3 à savoir que :

$$g \leq \frac{\sqrt{x_5} - \sqrt{x_1}}{\sqrt{x_5} - \sqrt{x_1}} = \frac{\sqrt{\frac{x_5}{x_1}} - 1}{\sqrt{\frac{x_5}{x_1}} + 1} = \frac{\sqrt{300} - 1}{\sqrt{300 + 1}} \simeq 0,89.$$

6.4 Exemples 5 et 6

Ces deux exemples sont théoriques : ils ont pour but de réillustrer, dans un cas où $p \neq 2,$ le fait que le coefficient de Gini ne caractérise pas, d'un point de vue concentration, une série.

Exemple 5

x_i	n_i	α_i	β_i	$\alpha_i - \beta_i$
$\frac{9,5}{20}$	20	0,2	0,095	0,105
$\frac{11,5}{20}$	20	0,4	0,21	0,19
$\frac{13,5}{20}$	20	0,6	0,345	0,255
$\frac{15,5}{20}$	20	0,8	0,5	0,3
$\frac{50}{20}$	20	1	1	0

Exemple 6

x_i	n_i	α_i	β_i	$\alpha_i - \beta_i$
$\frac{1}{20}$	20	0,2	0,01	0,19
$\frac{9}{20}$	20	0,4	0,1	0,3
$\frac{29}{20}$	20	0,6	0,39	0,21
$\frac{30}{20}$	20	0,8	0,69	0,11
$\frac{31}{20}$	20	1	1	0

Pour les deux exemples on a $p = 5$ et la masse totale est $m = \sum_{i=1}^{5} n_i x_i = 100.$

La lecture des dernières colonnes de chacun des tableaux prouve que ces deux séries ont le même coefficient $e = 0,3$ et, toujours par application de la formule

$g = 1 - \sum_{i=0}^{4} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1})$ on obtient (je laisse au lecteur le soin de le vérifier) $g_5 = 0,34$ et $g_6 = 0,324$

Les coefficients de Gini sont donc pratiquement égaux.

Et pourtant les courbes de Lorentz de ces deux séries sont trés différentes :

Essayons une analyse plus précise :

a) ***si on considére les 50% derniers individus***

pour l'exemple 4, ils possédent $\frac{\frac{13,5}{2}+15,5+50}{100} = 72\%$ de la masse totale.

pour l'exemple 5, ils possédent $\frac{\frac{29}{2}+30+31}{100} = 76\%$ de la masse totale.

A ce niveau la situation est à peu prés semblable : on a traduit ici le fait que les 2 courbes de Lorentz ont à peu prés la même ordonnée pour l'abcisse 0,5.

b) ***si on considére les 10% derniers individus***

pour l'exemple 4, ils possédent 25% de la masse totale

pour l'exemple 5, ils possédent 15,5% de la masse totale.

A ce niveau la situation est trés différente.

Pourquoi ne pas dire qu'au niveau des 50% derniers individus les deux séries ont la même concentration, mais la série 5 est plus concentrée à la fin que la série 6?

Cette conclusion s'obtient grâce uniquement au calcul (simple) de la masse possédée par les 50% derniers individus et de celle possédée par les 10% derniers individus et, évidemment le seul calcul (compliqué) de g ne permet pas d'arriver à une telle conclusion puisque pour les deux séries ont le même g!

Une fois de plus disons que le but des chapitres suivants sera de justifier que pour analyser en terme de concentration une série il est plus pertinent (et plus simple) de calculer la masse des 50% derniers individus et celle des 10% derniers individus.

6.5 Exemple 7

Il s'agit encore d'un exemple théorique correspondant à la situation suivante : les effectifs sont constants et la suite des valeurs prises par le caractère est une suite arithmétique croissante, c'est à dire $x_i = ia + b$ pour $i = 1, 2, .., p$ avec $b \geq 0$ et $a > 0$.

On a de façon immédiate, pour $k = 1, 2, \ldots, p$:

$$n_k = \frac{n}{p},\ \alpha_k = \frac{k}{p},\ \beta_k = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{p} n_i x_i} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{p} x_i} = \frac{kb + \frac{k(k+1)}{2}a}{S} \text{ avec } S = pb + \frac{p(p+1)}{2}a.$$

On constate tout de suite que les points $M_k(\alpha_k, \beta_k)$ sont sur la parabole d'équation $y = \frac{1}{S}(pbx + \frac{pax(px+1)}{2})$, parabole passant par $O(0,0)$ et $A(1,1)$, mais cette parabole n'est pas pour autant la courbe de Lorentz. Cependant pour p grand ces deux courbes seront pratiquement confondues.

Calcul du coefficient de Gini :
D'aprés 3.1.4

$$\begin{aligned} g &= 1 - \sum_{i=0}^{p-1} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1}) \\ &= 1 - \frac{1}{p}\sum_{i=0}^{p-1}(\beta_i + \beta_{i+1}) \\ &= 1 - \frac{1}{p}(\beta_0 + 2\beta_1 + \ldots + 2\beta_{p-1} + \beta_p) \\ &= 1 - \frac{1}{p} - \frac{2}{p}(\beta_1 + \ldots + \beta_{p-1}) \end{aligned}$$

Mais $\beta_k = \frac{1}{S}(k(b + \frac{a}{2}) + k^2\frac{a}{2})$ d'où $\beta_1 + \ldots + \beta_{p-1} = \frac{1}{S}(\frac{(p-1)p}{2}(b + \frac{a}{2}) + \frac{(p-1)p(2p-1)}{6}a)$

et $g = 1 - \frac{1}{p} - \frac{p-1}{S}(b + \frac{a}{2} + \frac{2p-1}{6}a) = \frac{a(p^2-1)}{p(3a(p+1)+6b)} = \frac{1}{3} \times \frac{1 - \frac{1}{p^2}}{1 + \frac{a+2b}{pa}}$.

On en déduit tout de suite $\underline{\lim_{p\to+\infty} g = \frac{1}{3}}$; (voir annexe 4 pour l'effet sur g d'une augmentation de b).

Que peut-on conclure?
Cette valeur limite est nettement inférieure au milieu de l'intervalle $[0;1]$ dans lequel se trouve g et donc à ce titre on pourrait parler de ‹‹concentration›› en dessous de la moyenne, cela sans avoir aucune information précise sur les groupes ayant plus en masse qu'en effectif.

Calcul de $e = \max_{k\in\{0;1;\ldots;p\}} \alpha_k - \beta_k$

$\alpha_k - \beta_k = \frac{k}{p} - \frac{kb + \frac{k(k+1)}{2}a}{S} = \frac{f(k)}{pS}$ avec $f(x) = \frac{pa}{2}(px - x^2)$

La fonction f ayant un maximum pour $x = \frac{p}{2}$, on a deux possibilités :

soit p est pair et alors $e = \frac{f(\frac{p}{2})}{pS} = \frac{p^2a}{8S}$ et donc pour p pair $\lim_{p\to+\infty} e = \frac{1}{4}$ (puisque $\lim_{p\to+\infty} \frac{S}{p^2} = \frac{a}{2}$)

soit p est impair, dans ce cas la valeur maximum de $f(k)$ sera $f(\frac{p+1}{2}) = f(\frac{p-1}{2}) = \frac{p(p^2-1)a}{8}$ et alors $e = \frac{(p^2-1)a}{8S}$ et pour p impair $\lim_{p\to+\infty} e = \frac{1}{4}$.

$\underline{\text{Finalement } \lim_{p\to+\infty} e = \frac{1}{4}}$

Ce résultat est tout à fait cohérent avec l'inégalité $\frac{g}{2} \le e \le g$ et conduit à la même conclusion que précédemment, à savoir ‹‹concentration›› en dessous de la moyenne, mais avec ce renseignement supplémentaire que l'écart maximum entre masse et effectif d'un même groupe (constitué de premiers ou derniers individus) est jutemment 0,25.
Cependant on peut ajouter que pour p grand, c'est le groupe des 50% premiers qui réalise le plus grand écart effectif moins masse, égal à $e = 0,25$, puisque $\frac{p+1}{2}$ équivaut à $\frac{p}{2}$ (pour p grand) et $\alpha\frac{p}{2} = 50\%$.

Allons plus loin. Toujours pour p grand, puisque les 50% premiers ayant 25% de la masse, c'est que les 50% derniers ont 75% de la masse et donc leur *mse* est de $\frac{75}{50} = 1,5$: ce *mse* pouvant varier entre 1 et 2, pourquoi ne pas dire que la concentration est moyenne au niveau des 50% derniers?

Remarque 1 :

Si $p = 2$ alors $g = 1 - \frac{1}{2} - \frac{a+b}{3a+2b} = \frac{a}{2(3a+2b)} = \frac{x_2 - x_1}{2(x_1+x_2)} = \frac{\text{étendue}}{4\bar{x}}$.

On peut vérifier que cette quantité est bien inférieure à $\frac{\sqrt{x_2} - \sqrt{x_1}}{\sqrt{x_2} + \sqrt{x_1}}$ (voir 3.2.3).

Remarque 2 :

Si $b = 0$ alors $g = \frac{p-1}{3p} = \frac{1}{3} - \frac{1}{3p}$.

Remarque 3 :

La méthode Gini-Lorentz a été présentée ici dans le cas discret, mais cette méthode s'adapte au cas où les valeurs du caractère sont considérées comme les valeurs prises par une variable aléatoire continue. Par exemple si cette variable aléatoire suit une loi uniforme sur l'intervalle $[0; a]$ alors la courbe de Lorentz est la parabole d'équation $y = x^2$ et $g = \frac{1}{3}$ (voir la référence [3] page 637)

6.6 Conclusion

Ces exemples montrent clairement que les commentaires faits au 5.3 sont justifiés.

Le coefficient de Gini est lourd à calculer et n'apporte pas d'éclairage précis sur la série.

Par contre le coefficient *e*, immédiat à calculer, apporte lui une information très précise : c'est l'écart maximum entre effectif et masse pour tous les groupes constitués soit des premiers individus, soit des derniers individus, et le groupe réalisant cet écart maximum est facile à mettre en évidence. Cet écart peut varier entre 0 (répartition égalitaire) et 1 (beaucoup ont peu) et il ***permet donc exactement le même type de conclusion que le coefficient de Gini***.

A ce titre, et compte tenu donc de sa simplicité de calcul et de son interprétation économique très concrète, le coefficient *e* ***peut remplacer avantageusement le coefficient de Gini***.

Cependant l'examen des *mse* des 50% derniers individus et des 10% derniers individus semblent beaucoup plus pertinent encore à considérer : les chapitres suivants vont développer ce point de vue.

7 Compléments sur les rapports masse sur effectif (*mse*)

On a vu aux chapitres 4 et 5 que les coefficients g et e reposent sur la considération des différences $\alpha - m_{pr}(\alpha)$ ou $m_{dr}(\alpha) - \alpha$, $\frac{1}{2}g$ en étant la valeur moyenne et e la valeur maximum.

On va s'intéresser ici aux rapports $\frac{m_{pr}(\alpha)}{\alpha}$ et $\frac{m_{dr}(\alpha)}{\alpha}$ qui sont en fait respectivement les rapports masse sur effectif des groupes constitués par les α premiers individus et des α derniers individus.

La notion de rapport masse sur effectif a déjà été définie au chapitre 2 : il s'agit ici de généraliser et de donner des résultats complémentaires.

7.1 Définition

‖ Les individus étant classés par valeur croissante du caractère et α étant dans $[0;1]$
‖ (α représentera toujours un effectif exprimé en pourcentage de l'effectif total),
‖ on notera $G_{pr}(\alpha)$ le groupe des α premiers individus et
‖ on notera $G_{dr}(\alpha)$ le groupe des α derniers individus
‖ Théoriquement ces groupes n'existent réellement que si α est de la forme $\frac{k}{n}$ avec k entier
‖ compris entre 0 et n, k correspondant au nombre d'individus du groupe.
‖ Les masses possédées par $G_{pr}(\alpha)$ et $G_{dr}(\alpha)$ sont alors respectivement $m_{pr}(\alpha)$ et $m_{dr}(\alpha)$
‖ (voir 4.1, 4.2). Mais puisque les fonctions m_{pr} et m_{dr} sont définies sur $[0;1]$ (voir 4.1)
‖ on considérera ces groupes pour α dans $[0;1]$ avec, pour masse possédée $m_{pr}(\alpha)$ et $m_{dr}(\alpha)$.
‖ Ainsi: $\boxed{\forall \alpha \in]0;1] \quad mse(G_{pr}(\alpha)) = \frac{m_{pr}(\alpha)}{\alpha} \text{ et } mse(G_{dr}(\alpha)) = \frac{m_{dr}(\alpha)}{\alpha}}$

Remarque 1 :

$G_{pr}(0) = G_{dr}(0) = \emptyset$ et $G_{pr}(1) = G_{dr}(1) =$ le groupe constitué de tous les individus.

$G_{pr}(\alpha_i)$ est le groupe des individus ayant une valeur du caractère $\leq x_i$ (pour $i = 1,2,\ldots,p$)

$G_{dr}(1-\alpha_i)$ est le groupe des individus ayant une valeur du caractère $\geq x_{i+1}$ (pour $i = 0,1,\ldots,p-1$)

Remarque 2 :

On a vu au chapitre 2 que pour tout groupe G constitué de k individus (k étant un nombre entier) on a :

$$mse(G) = \frac{\text{moyenne du caractère pour le groupe } G}{\text{moyenne de la série}} = \frac{\bar{x}_G}{\bar{x}}$$

et donc les groupes ayant les plus forts *mse* (c'est à dire $\frac{x_p}{\bar{x}}$) sont les sous groupes des n_p derniers individus (ceux ayant la valeur x_p du caractère). On a un résultat analogue pour les plus faibles *mse*.

Remarque 3 :

On verra en annexe 4 l'influence sur les *mse* d'un ajout d'une même quantité à toutes les valeurs du caractère.

7.2 Propriété

1) ‖ Si $p = 1$ $\forall \alpha \in]0;1]$ $mse(G_{pr}(\alpha)) = mse(G_{dr}(\alpha)) = 1$
2) ‖ Pour $p \geq 2$ la fonction $mse(G_{pr}(\alpha))$ est constante ($\frac{x_1}{\bar{x}}$) sur $]0;\alpha_1]$ et strictement croissante sur
‖ $[\alpha_1;1]$, donc croissante sur $]0;1]$.
3) ‖ Pour $p \geq 2$ la fonction $mse(G_{dr}(\alpha))$ est constante ($\frac{x_p}{\bar{x}}$) sur $]0;1-\alpha_{p-1}]$
‖ et strictement décroissante sur $[1-\alpha_{p-1};1]$, donc décroissante sur $]0;1]$.

Preuve :

1) Evident (tout *mse* est égal à 1, voir 1 de 2.2.3).

2) D'après l'annexe 2 pour $i = 0,1,\ldots,p-1,$ sur $[\alpha_i;\alpha_{i+1}]$ $m_{pr}(\alpha) = a_i\alpha + b_i$ avec $a_i = \frac{x_{i+1}}{\bar{x}}$ et $b_0 = 0$ et $b_i < 0$ pour $1 \leq i \leq p-1$.
Donc sur $]0;\alpha_1]$ $\frac{m_{pr}(\alpha)}{\alpha} = a_0 = \frac{x_1}{\bar{x}}$ et pour $1 \leq i \leq p-1$ sur $[\alpha_i;\alpha_{i+1}]$ $\frac{m_{pr}(\alpha)}{\alpha} = a_i + \frac{b_i}{\alpha}$, expression qui est strictement croissante sur $[\alpha_i;\alpha_{i+1}]$ puisque $b_i < 0$.
La fonction $\frac{m_{pr}(\alpha)}{\alpha}$ étant continue sur $]0;1]$ on peut donc dire qu'elle est strictement croissante sur $[\alpha_1;1]$.

3) D'après 4.2 $m_{dr}(\alpha) = 1 - m_{pr}(1-\alpha)$ et donc pour $i = 0,1,\ldots,p-1,$ sur $[1-\alpha_{i+1};1-\alpha_i]$ on a $1-\alpha \in [\alpha_i;\alpha_{i+1}]$ ce qui donne $m_{dr}(\alpha) = 1 - (a_i(1-\alpha) + b_i) = a_i\alpha + 1 - a_i - b_i$.
Donc sur $]0;1-\alpha_{p-1}]$ $\frac{m_{dr}(\alpha)}{\alpha} = a_{p-1} + \frac{1-a_{p-1}-b_{p-1}}{\alpha} = a_{p-1}$ (voir annexe 2 pour $1-a_{p-1}-b_{p-1} = 0$).
Pour $0 \leq i \leq p-2$ sur $[1-\alpha_{i+1};1-\alpha_i]$ on a $\frac{m_{dr}(\alpha)}{\alpha} = a_i + \frac{1-a_i-b_i}{\alpha}$ expression qui est strictement décroissante sur $[1-\alpha_{i+1};1-\alpha_i]$ puisque $1-a_i-b_i > 0$ (voir annexe 2).
La fonction $\frac{m_{dr}(\alpha)}{\alpha}$ étant continue sur $]0;1]$ on peut donc dire qu'elle est strictement décroissante sur $[1-\alpha_{p-1};1]$.

Remarque :
On verra à l'annexe 6 un exemple de représentation graphique de la fonction $\alpha \mapsto mse(G_{dr}(\alpha))$

7.3 Propriété

‖ Si $0 < \alpha_2 < \alpha_1 < 1$ alors
1) ‖ $mse(G_{dr}(\alpha_1)) = mse(G_{dr}(\alpha_2))$ $\Leftrightarrow$la répartition est égalitaire à l'intérieur du groupe $G_{dr}(\alpha_1)$
2) ‖ $mse(G_{pr}(\alpha_1)) = mse(G_{pr}(\alpha_2))$ $\Leftrightarrow$la répartition est égalitaire à l'intérieur du groupe $G_{pr}(\alpha_1)$

Preuve :

1) Si $mse(G_{dr}(\alpha_1) = mse(G_{dr}(\alpha_2))$ on a $\frac{m_{dr}(\alpha_2)}{m_{dr}(\alpha_1)} = \frac{\alpha_2}{\alpha_1}$. Considérons alors la série *Se* constituée des α_1 derniers individus de la série initiale. Elle a pour masse totale $m' = m_{dr}(\alpha_1)m$ et les $\frac{\alpha_2}{\alpha_1}$ (en pourcentage) derniers individus de cette série *Se* (qui correspondent aux α_2 derniers individus de la série initiale) possédent alors (en pourcentage de la masse totale de cette série *Se*) $\frac{m_{dr}(\alpha_2)m}{m'} = \frac{\alpha_2}{\alpha_1}$ qui est dans $]0;1[$. Donc d'après 4.3, appliqué à la série *Se,* il y répartition égalitaire pour cette série c'est-à-dire pour le groupe $G_{dr}(\alpha_1)$.
Dans l'autre sens c'est immédiat.

2) Démonstration analogue.

.

7.4 Propriété

1) ‖ $\forall\alpha \in]0;1]$ $\frac{x_1}{\bar{x}} \leq mse(G_{pr}(\alpha)) \leq 1 \leq mes(G_{dr}(\alpha)) \leq \frac{x_p}{\bar{x}}$
2) ‖ C_{pr} étant la courbe de Lorentz, représentation graphique de la fonction m_{pr},
‖ $\forall\alpha \in]0;1]$ on a :
‖ $mse(G_{pr}(\alpha))$ = pente de la droite (ON_α) avec N_α point d'abcisse α de C_{pr}.
‖ $mse(G_{dr}(\alpha))$ = pente de la droite $(AN_{1-\alpha})$ avec $N_{1-\alpha}$ point d'abcisse $1-\alpha$ de C_{pr}.
3) ‖ Si $p \geq 2$:
‖ $mse(G_{pr}(\alpha_k)) = \frac{\beta_k}{\alpha_k}$ = pente (≤ 1)de la droite (OM_k) , cela pour $k = 1,2,\ldots,p$
‖ la suite $(\frac{\beta_k}{\alpha_k})_{k\in\{1,2,\ldots,p\}}$ est strictement croissante :

‖ elle commence à $\frac{\beta_1}{\alpha_1} = \frac{x_1}{\bar{x}}$ (le plus petit *mse*) et elle finit à $\frac{\beta_p}{\alpha_p} = 1$

‖ $mse(G_{dr}(1-\alpha_k)) = \frac{1-\beta_k}{1-\alpha_k}$ = pente (≥ 1)de la droite (AM_k) cela pour $k = 0, 1, \ldots, p-1$

‖ la suite $(\frac{1-\beta_k}{1-\alpha_k})_{k\in\{0,1,\ldots,p-1\}}$ est strictement croissante :

‖ elle commence à $\frac{1-\beta_0}{1-\alpha_0} = 1$ et elle finit à $\frac{1-\beta_{p-1}}{1-\alpha_{p-1}} = \frac{x_p}{\bar{x}}$ (le plus grand *mse*)

‖ Rappel : les points M_k sont les points de coordonnées (α_k, β_k) (voir 3.1.2) et donc $M_k = N_{\alpha_k}$.

Preuve :

1) Si $p = 1$ c'est évident. Supposons $p \geq 2$.

Par définition $mse(G_{pr}(\alpha)) = \frac{m_{pr}(\alpha)}{\alpha}$ et $mse(G_{dr}(\alpha)) = \frac{m_{dr}(\alpha)}{\alpha}$ or d'après 4.3 on a $m_{pr}(\alpha) \leq \alpha \leq m_{dr}(\alpha)$ ce qui donne la partie centrale des inégalités en divisant par α.

Pour α de la forme $\frac{k}{n}$ avec k entier (de 0 à n) les inégalités extrêmes résultent de 2.2.3.

Une preuve pour α quelconque dans $]0;1]$ consiste à utiliser 7.2 : en effet l'expression $mse(G_{pr}(\alpha))$ est croissante sur $]0;1]$ or elle est égale à $\frac{x_1}{\bar{x}}$ pour $\alpha \in]0;\alpha_1]$ donc sa valeur ***minimale*** est $\frac{x_1}{\bar{x}}$.

Même raisonnement pour $mse(G_{dr}(\alpha))$.

2) N_α a pour coordonnées $(\alpha, m_{pr}(\alpha))$ donc la pente la droite (ON_α) est $\frac{m_{pr}(\alpha)-0}{\alpha-0} = mse(G_{pr}(\alpha))$.

$N_{1-\alpha}$ a pour coordonnées $(1-\alpha, m_{pr}(1-\alpha)) = (1-\alpha, 1-m_{dr}(\alpha))$ et A a pour coordonnée $(1,1)$ et donc la pente de la droite $(AN_{1-\alpha})$ est $\frac{1-m_{dr}(\alpha)-1}{1-\alpha-1} = mse(G_{dr}(\alpha))$.

3) C'est une application immédiate du point 2 précedent :

$mse(G_{pr}(\alpha_k))$ = pente de la droite $(ON_{\alpha_k}) = (OM_k)$ et par ailleurs ce *mse* est $\frac{m_{pr}(\alpha_k)}{\alpha_k} = \frac{\beta_k}{\alpha_k}$.

La stricte croissance de cette suite résulte de 7.2 (la fonction $\alpha \mapsto mse(G_{pr}(\alpha))$ est strictement croissante sur $[\alpha_1;1]$ et la suite $(\alpha_k)_{k\in\{1;2;\ldots;p\}}$ est évidemment strictement croissante).

On verra en annexe 1 une 2ième preuve sans utiliser les *mse*.

On peut en donner une 3ième preuve en terme de moyenne : lorsqu'on passe du groupe $G_{pr}(\alpha_k)$ au groupe $G_{pr}(\alpha_{k+1})$ on ajoute tous les individus ayant la valeur x_{k+1} du caractère et comme cette valeur est strictement supérieure à toutes les valeurs du caractère des individus de $G_{pr}(\alpha_k)$ on augmente la moyenne, c'est à dire $\bar{x}_{G_{pr}(\alpha_k)} \leq \bar{x}_{G_{pr}(\alpha_{k+1})}$ et en divisant par $\bar{x}$ on obtient $mse(G_{pr}(\alpha_k)) \leq mse(G_{pr}(\alpha_{k+1}))$.

Considérons maintenant la suite $mse(G_{dr}(1-\alpha_k))$.

$mse(G_{dr}(1-\alpha_k))$ = pente de la droite $(AN_{\alpha_k}) = (AM_k)$ et par ailleurs ce *mse* est $\frac{m_{dr}(1-\alpha_k)}{1-\alpha_k} = \frac{1-m_{pr}(\alpha_k)}{1-\alpha_k} = \frac{1-\beta_k}{1-\alpha_k}$.

La stricte croissance de cette suite résulte de 7.2 (la fonction $\alpha \mapsto mse(G_{dr}(\alpha))$ est strictement décroissante sur $[1-\alpha_{p-1};1]$, donc $mse(G_{dr}(1-\alpha))$ est strictement croissante sur $[0;\alpha_{p-1}]$ et la suite α_k ($k = 0$ à $k = p-1$) est strictement croissante).

On verra en annexe 1 une 2ième preuve sans utiliser les *mse*, une 3ième preuve pouvant être faite en terme de moyenne comme ci-dessus.

Remarque 1 :

La double inégalité $mse(G_{pr}(\alpha)) \leq 1 \leq mes(G_{dr}(\alpha))$ traduit le fait que $G_{dr}(\alpha)$ contient les individus ayant les plus fortes valeurs du caractère et que $G_{pr}(\alpha)$ contient les individus ayant les plus faibles valeurs du caractère.

Lorsque α est de la forme $\frac{k}{n}$ avec k entier (de 0 à n) on peut faire une démonstration de cette double inégalité en termes de moyenne : pour passer du groupe $G_{dr}(\alpha)$ au groupe constitué de tous les individus on ajoute des individus dont les valeurs du caractère sont toutes inférieures ou égales à celles des individus du groupe $G_{dr}(\alpha)$ donc $\bar{x} \leq \bar{x}_{G_{dr}(\alpha)}$, et en divisant par $\bar{x}$ on obtient $1 \leq \frac{\bar{x}_{G_{dr}(\alpha)}}{\bar{x}} = mse(G_{dr}(\alpha))$.

Remarque 2 :

La stricte croissance de la suite $\frac{\alpha_k}{\beta_k}$ a déjà été signalée au 3.1.3

Remarque 3 :
Le résultat 3 ci-dessus sera traduit de façon pratique au 7.6

7.5 Propriété

|| Si $\alpha = \frac{k}{n}$ avec $k \in \{1; 2; \ldots; n\}$ et si G est un groupe (constitué d'un nombre entier
|| d'individus) d'effectif (en pourcentage) ***supérieur ou égal*** à α, alors :
|| $mse(G_{pr}(\alpha)) \leq mse(G) \leq mse(G_{dr}(\alpha))$.
|| On notera que les individus du groupe G ne se ‹‹suivent›› pas forcément au point de vue valeur
|| du caractère.

Preuve :

Notons t le nombre d'individus du groupe G : $k \leq t \leq n$
1er cas
$G \subset G_{dr}(\alpha)$, donc obligatoirement $t = k$ et $G = G_{dr}(\alpha)$ et on a $\bar{x}_G \leq \bar{x}_{G_{dr}(\alpha)}$
2iéme cas
$G \cap G_{dr}(\alpha) = \emptyset$, c'est-à-dire les t individus de G sont ‹‹avant›› ceux de de $G_{dr}(\alpha)$ et donc ils ont tous une valeur du caractère inférieure ou égale à la plus petite valeur du caractère des individus de $G_{dr}(\alpha)$: on a encore $\bar{x}_G \leq \bar{x}_{G_{dr}(\alpha)}$.
3iéme cas
q individus de G sont dans $G_{dr}(\alpha)$, les $t - q$ autres étant ‹‹avant›› avec $0 < q \leq k$ et $q < t$.
Posons $r = t - k \geq 0$: les t individus de G se répartissent ainsi (en les classant toujours par valeur croissante du caractère) :

r individus ‹‹avant›› $G_{dr}(\alpha)$: ils constituent le groupe G_1
puis $k - q$ individus situés encore ‹‹avant›› $G_{dr}(\alpha)$: ils constituent le groupe G_2
puis les q individus situés dans $G_{dr}(\alpha)$: ils constituent le groupe G_3

Notons que l'on peut avoir dans ce 3iéme cas $G_1 = \emptyset$ (G et $G_{dr}(\alpha)$ ont le même effectif) ainsi que $G_2 = \emptyset$ (G contient $G_{dr}(\alpha)$), mais G_1 et G_2 ne peuvent être vides simultanément.
Si $r = 0$ alors $G_1 = \emptyset$ et $G = G_2 \cup G_3$ et $\bar{x}_G = \bar{x}_{G_2 \cup G_3}$ sinon, on passe de $G_2 \cup G_3$ à $G = G_1 \cup G_2 \cup G_3$ en ajoutant à $G_2 \cup G_3$ les r individus de G_1 qui ont tous une valeur du caractère inférieure ou égale à la plus petite valeur du caractère des individus de $G_2 \cup G_3$, et donc on diminue là aussi la moyenne : $\bar{x}_G \leq \bar{x}_{G_2 \cup G_3}$.
Si $q = k$ alors $G_2 = \emptyset$, $G_d(\alpha) = G_3 = G_2 \cup G_3$ et $\bar{x}_{G_2 \cup G_3} = \bar{x}_{G_{dr}(\alpha)}$ sinon, on passe de $G_{dr}(\alpha)$ à $G_2 \cup G_3$ en remplacant les $k - q$ individus de $G_{dr}(\alpha)$ n'appartenant à G_3 par les $k - q$ individus de G_2, ce qui revient à diminuer la valeur du caractère de chacun de ces $k - q$ individus de $G_{dr}(\alpha)$ (puisque G_2 est ‹‹avant›› $G_{dr}(\alpha)$), et donc on diminue la moyenne : $\bar{x}_{G_2 \cup G_3} \leq \bar{x}_{G_{dr}(\alpha)}$.
Dans ce 3iéme cas on a donc toujours $\bar{x}_G \leq \bar{x}_{G_2 \cup G_3}$ et $\bar{x}_{G_2 \cup G_3} \leq \bar{x}_{G_{dr}(\alpha)}$ soit $\bar{x}_G \leq \bar{x}_{G_{dr}(\alpha)}$.

Finalement dans les 3 cas on a $\bar{x}_G \leq \bar{x}_{G_{dr}(\alpha)}$ et en divisant par $\bar{x}$ on obtient $mse(G) \leq mse(G_{dr}(\alpha))$.
Même méthode pour démontrer $mse(G_{pr}(\alpha)) \leq mse(G)$.

Remarque :
On peut retrouver (partiellement du moins) la propriété 7.2 à partir de 7.5.
Si $e_1 \leq e_2$ (avec e_1 et e_2 de la forme $\frac{k}{n}$ et k entier entre 1 et n)
alors $G_{dr}(e_2)$ est un groupe d'effectif $\geq e_1$ et donc $mse(G_{dr}(e_2)) \leq mse(G_{dr}(e_1))$
et $G_{pr}(e_2)$ est un groupe d'effectif $\geq e_1$ et donc $mse(G_{pr}(e_1)) \leq mse(G_{pr}(e_2))$.
Ce qui correspond à la décroissance de la fonction $\alpha \mapsto mse(G_{dr}(\alpha)) = \frac{m_{dr}(\alpha)}{\alpha}$ et à la croissance de la fonction $\alpha \mapsto mse(G_{pr}(\alpha)) = \frac{m_{pr}(\alpha)}{\alpha}$, mais en se limitant ici à des valeurs de la forme $\frac{k}{n}$ pour la variable α.

7.6 Propriété

‖ rappel (voir 2.1) : $p \geq 1$; $0 < x_1 < x_2 \ldots < x_p$; $n_1, n_2, \ldots, n_p > 0$

1) ‖ Il a été vu au 7.4 que $\frac{x_p}{\bar{x}}$ est la valeur maximum de tous les *mse*, en particulier de tous les
‖ $mse(G_{dr}(\alpha))$ pour $\alpha \in]0;1]$; cette valeur maximum dépend de la valeur moyenne, donc des
‖ effectifs. D'après 2.2.3 $mse(G_{dr}(\alpha))$ sera égal à $\frac{x_p}{\bar{x}}$ si et seulement si $G_{dr}(\alpha)$ est un
‖ sous-groupe des n_p derniers individus, c'est-à-dire : $mse(G_{dr}(\alpha)) = \frac{x_p}{\bar{x}} \Leftrightarrow \alpha \leq \frac{n_p}{n}$.
‖ ***Donnons deux autres majorations***

2) ‖ $\forall \alpha \in]0;1[$ on a $1 \leq mse(G_{dr}(\alpha)) \leq \frac{x_p}{\alpha x_p + (1-\alpha)x_1} < \frac{1}{\alpha}$
‖ on a donc ici pour α fixé dans $]0;1[$ deux autres majorants de $mse(G_{dr}(\alpha))$:
‖ $\frac{x_p}{\alpha x_p + (1-\alpha)x_1}$ est un majorant valable pour toutes les séries à x_1 à x_p fixés
‖
‖ $\frac{1}{\alpha}$ est un majorant pour toutes les séries.

3) ‖ Pour $\alpha \in]0;1[$ l'inégalité ***stricte*** $mse(G_{dr}(\alpha)) < \frac{1}{\alpha}$ est en fait une conséquence immédiate
‖ de l'inégalité $m_{dr}(\alpha) < 1$ laquelle découle de $x_1 > 0$ (car alors les premiers ont forcément
‖ quelque chose et les derniers ne peuvent avoir tout).
‖ Par contre si on suppose $x_1 = 0$ (et $p \geq 2$) alors
‖ $mse(G_{dr}(\alpha)) = \frac{1}{\alpha} \Leftrightarrow$ les $1-\alpha$ premiers ont 0 (donc les α derniers ont tout)
‖ cette situation n'exigeant pas que $p = 2$.

4) ‖ Pour $\alpha \in]0;1[$:
‖ $mse(G_{dr}(\alpha)) = \frac{x_p}{\alpha x_p + (1-\alpha)x_1} \Leftrightarrow p = 2, \frac{n_1}{n} = 1-\alpha, \frac{n_2}{n} = \alpha$
‖ Pour ce résultat, l'hypothése $x_1 > 0$ est capitale : on n'a pas le droit ici de faire $x_1 = 0$,
‖ d'ailleurs on arriverait à une contradiction avec le 3 précédent.

5) ‖ Pour $\alpha \in]0;1[$:
‖ sur l'ensemble de toutes les séries $mse(G_{dr}(\alpha))$ décrit tout l'intervalle $\left[1; \frac{1}{\alpha}\right[$

Preuve :

1) Faite au cours de l'énoncé.

2) La minoration par 1 est une redite de 7.4. Prouvons la majoration.

$$m_{dr}(\alpha) = \frac{m_{dr}(\alpha)}{m_{dr}(\alpha) + m_{pr}(1-\alpha)} = \frac{1}{1 + \frac{m_{pr}(1-\alpha)}{m_{dr}(\alpha)}} \quad (\alpha > 0 \text{ et } x_1 > 0 \text{ assurent que } m_{dr}(\alpha) > 0)$$

De 7.4 on tire $mse(G_{pr}(1-\alpha)) \geq \frac{x_1}{\bar{x}}$ et $mse(G_{dr}(\alpha)) \leq \frac{x_p}{\bar{x}}$ soit

$$m_{pr}(1-\alpha) \geq \frac{x_1}{\bar{x}}(1-\alpha) \text{ et } \frac{1}{m_{dr}(\alpha)} \geq \frac{\bar{x}}{\alpha x_p}$$

Par multiplication membre à membre on obtient

$\frac{m_{pr}(1-\alpha)}{m_{dr}(\alpha)} \geq \frac{(1-\alpha)x_1}{\alpha x_p}$, l'égalité ayant lieu si et seulement si $m_{pr}(1-\alpha) = \frac{x_1}{\bar{x}}(1-\alpha)$ et $\frac{1}{m_{dr}(\alpha)} = \frac{\bar{x}}{\alpha x_p}$.

Finalement $m_{dr}(\alpha) \leq \frac{1}{1 + \frac{(1-\alpha)x_1}{\alpha x_p}} = \frac{\alpha x_p}{\alpha x_p + (1-\alpha)x_1}$ et en divisant par α on obtient la majoration annoncée :

$mse(G_{dr}(\alpha)) \leq \frac{x_p}{\alpha x_p + (1-\alpha)x_1}$.

3) Faite au cours de l'énoncé.

4) Compte tenu du calcul fait au cours de la preuve du point 2 ci-dessus :

$mse(G_{dr}(\alpha)) = \frac{x_p}{\alpha x_p + (1-\alpha)x_1} \Leftrightarrow \frac{m_{pr}(1-\alpha)}{m_{dr}(\alpha)} = \frac{(1-\alpha)x_1}{\alpha x_p}$

$\Leftrightarrow m_{pr}(1-\alpha) = \frac{x_1}{\bar{x}}(1-\alpha)$ et $\frac{1}{m_{dr}(\alpha)} = \frac{\bar{x}}{\alpha x_p}$

$\Leftrightarrow mse(G_{pr}(1-\alpha)) = \frac{x_1}{\bar{x}}$ et $mse(G_{dr}(\alpha)) = \frac{x_p}{\bar{x}}$.

D'aprés 2.2.3 $G_{pr}(1-\alpha)$ est alors un sous-groupe des n_1 premiers individus et $G_{dr}(\alpha)$ un sous-groupe des n_p derniers individus d'où $1-\alpha \leq \frac{n_1}{n}$ et $\alpha \leq \frac{n_p}{n}$ ce qui entraîne $n \leq n_1 + n_p$; mais $n = n_1 + n_2 + \ldots + n_p$ et les n_i sont strictement positifs, donc la seule possibilité est $p = 2$ et dans ce cas $1-\alpha \leq \frac{n_1}{n}$ donne $\alpha \geq \frac{n_2}{n}$ d'où $\alpha = \frac{n_2}{n}$ et $1-\alpha = \frac{n_1}{n}$.

La réciproque est immédiate à vérifier :

si $p = 2$, et si $1-\alpha = \frac{n_1}{n}$ (donc $\alpha = \frac{n_2}{n}$) alors

$mse(G_{dr}(\alpha)) = \frac{\frac{n\alpha x_2}{n(1-\alpha)x_1 + n\alpha x_2}}{\alpha} = \frac{x_2}{\alpha x_2 + (1-\alpha)x_1}$.

5) Si $p = 1$, $mse(G_{dr}(\alpha)) = 1$.

Si $p = 2$, en considérant une série telle que $1-\alpha = \frac{n_1}{n}$ on a :

$mse(G_{dr}(\alpha)) = \frac{x_2}{\alpha x_2 + (1-\alpha)x_1} = \frac{1}{\alpha + (1-\alpha)\frac{x_1}{x_2}}$.

Comme $0 < x_1 < x_2$, sur l'ensemble de toutes les séries $\frac{x_1}{x_2}$ décrit tout $]0;1[$ et dans ce cas l'expression $\frac{1}{\alpha + (1-\alpha)\frac{x_1}{x_2}}$ décrit tout l'intervalle $\left]1; \frac{1}{\alpha}\right[$.

Remarque 1 :

On peut donner une explication intuitive de l'inégalité $mse(G_{dr}(\alpha)) \leq \frac{x_p}{\alpha x_p + (1-\alpha)x_1}$, équivalente à $m_{dr}(\alpha) \leq \frac{\alpha x_p}{\alpha x_p + (1-\alpha)x_1}$. En effet à x_1 et x_p fixés, $m_{dr}(\alpha)$ (qui est la masse possédée par les α derniers individus) sera maximum si et seulement si les α derniers individus possédent le plus possible, c'est-à-dire s'ils ont tous x_p comme valeur du caractère et si la masse totale est la plus faible possible, c'est-à-dire si les $(1-\alpha)$ autres individus ont x_1 comme valeur du caractère, ce qui donne alors $m_{dr}(\alpha) = \frac{\alpha x_p}{\alpha x_p + (1-\alpha)x_1}$.

Remarque 2 :

Si $p = 1$ alors $\frac{x_p}{\bar{x}} = \frac{x_p}{\alpha x_p + (1-\alpha)x_1} = 1$

Si $p \geq 2$, la série étant fixée, alors selon les valeurs de α dans $]0;1[$, $\frac{x_p}{\bar{x}}$ peut être inférieur, ou égal, ou supérieur à $\frac{x_p}{\alpha x_p + (1-\alpha)x_1}$ et à $\frac{1}{\alpha}$. En effet $\frac{x_p}{\bar{x}}$ est dans $\left]1; \frac{x_p}{x_1}\right[$ et lorsque α décrit $]0;1[$, $\frac{x_p}{\alpha x_p + (1-\alpha)x_1}$ décrit tout $\left]1; \frac{x_p}{x_1}\right[$ et $\frac{1}{\alpha}$ décrit tout $]1;+\infty[$

Si $mse(G_{dr}(\alpha)) = \frac{x_p}{\bar{x}}$ alors $\frac{x_p}{\bar{x}} \leq \frac{x_p}{\alpha x_p + (1-\alpha)x_1}$ (d'après le 2 de la propriété)

Si $mse(G_{dr}(\alpha)) = \frac{x_p}{\alpha x_p + (1-\alpha)x_1}$ alors $\frac{x_p}{\bar{x}} = \frac{x_p}{\alpha x_p + (1-\alpha)x_1}$ (car d'après le 4 on a $p = 2$, $1-\alpha = \frac{n_1}{n}, \alpha = \frac{n_2}{n}$ et donc $\bar{x} = \alpha x_p + (1-\alpha)x_1$)

7.7
Obtention pratique des ‹‹principaux›› groupes ayant un $mse \leq 1$ et des ‹‹principaux›› groupes ayant un $mse \geq 1$

On va traduire ici de façon pratique les résultats du 3 de 7.4 : les principaux groupes ayant un $mse \leq 1$ sont les $G_{pr}(\alpha_k)$ et les principaux groupes ayant un $mse \geq 1$ sont les $G_{dr}(1-\alpha_k)$.

On a $mse(G_{pr}(\alpha_k)) = \frac{\beta_k}{\alpha_k}$ pour $k = 1, 2, \ldots, p$ et $mse(G_{dr}(1-\alpha_k)) = \frac{1-\beta_k}{1-\alpha_k}$ pour $k = 0, 1, \ldots, p-1$.

Ils sont faciles à mettre en évidence en remarquant que α_k et β_k sont les effectifs et masses en cumulés croissants (en pourcentage) alors que $1-\alpha_k$ et $1-\beta_k$ sont les effectifs et masses en cumulés décroissants (en pourcentage).

Illustrons cela sur l'exemple 1 du 6.1

			$\Sigma \frac{n_i x_i}{m} \nearrow$	$\Sigma \frac{n_i}{n} \nearrow$	$mse(G_{pr}(\alpha_i))$	$\Sigma \frac{n_i x_i}{m} \searrow$	$\Sigma \frac{n_i}{n} \searrow$	$mse(G_{dr}(1-\alpha_{i-1}))$
x_i	n_i	$n_i x_i$	β_i	α_i	$\frac{\beta_i}{\alpha_i}$	$1-\beta_{i-1}$	$1-\alpha_{i-1}$	$\frac{1-\beta_{i-1}}{1-\alpha_{i-1}}$
5	16	80	0,0375	0,2	0,1875	1	1	1
15	30	450	0,248	0,575	0,431	0,962	0,8	1,2
30	18	540	0,502	0,8	0,6275	0,752	0,425	1,77
55	10	550	0,760	0,925	0,82	0,498	0,2	2,49
85	6	510	1	1	1	0,24	0,075	3,19
	$n = 80$	$m = 2130$						

Rappel : $\alpha_0 = \beta_0 = 0$

Analyse du tableau

1) On vérifie bien que les suites $\frac{\beta_i}{\alpha_i}$ et $\frac{1-\beta_{i-1}}{1-\alpha_{i-1}}$ sont croissantes (pour $i = 1, 2, \ldots, p$)

2) Le plus petis *mse* est 0,1875 réalisé par les 20% premiers individus ($0,1875 = \frac{x_1}{\bar{x}}$) ; ce groupe a donc 5 fois moins en masse qu'en effectif.

3) Le plus fort *mse* est 3,19 réalisé par les 7,5% derniers individus ($3,19 = \frac{x_p}{\bar{x}}$)

4) Le groupe d'individus ayant une valeur du caractère ≥ 55 (les derniers 20% d'individus) a presque 2,5 fois plus en masse qu'en effectif, par contre si on descend au groupe des individus ayant une valeur du caractère ≥ 30 (les derniers 42,5%) celui ci n'a plus que 1,77 comme rapport masse sur effectif.

5) Notons que la propriété 7.5 permet d'affirmer que TOUT groupe d'individus d'effectif supérieur ou égal à 42,5% aura un *mse* inférieur ou égal à 1,77.

6) Quant à la propriété 7.2 elle permet d'affirmer que le groupe $G_{pr}(30\%)$ a un *mse* compris entre 0,1875 et 0,431 et que le groupe $G_{dr}(30\%)$ a un *mse* compris entre 1,77 et 2,485.

Conclusion :

On dispose ici d'un moyen simple (il suffit de calculer les effectifs et masses en cumulés croissants et décroissants et de faire quelques divisions) permettant de mettre en évidence tous les ‹‹principaux›› groupes ayant plus en masse qu'en effectif et ceux ayant moins en masse qu'en effectif.

Cependant il n'est pas facile d'en tirer une conclusion précise en terme de concentration, d'autant plus que lorsque l'on passe d'une série à une autre, les effectifs de ces ‹‹principaux›› groupes changent.

Le chapitre suivant va montrer que pour atteindre ce but il suffit de calculer deux *mse* bien choisis, quitte ensuite (si on le désire) à compléter ce résultat par la liste des principaux groupes évoqués ci-dessus, mais ce ne sera pas indispensable.

8 Recherche d'une nouvelle méthode d'analyse de la concentration d'une série.

8.1 UNE FAMILLE D'INDICATEURS DE CONCENTRATION.

Soit α fixé dans $]0;1[$: la propriété 4.3 prouve que $m_{dr}(\alpha)$ (masse, en pourcentage, possédée par les α derniers individus) est toujours supérieur ou égal à α et que $m_{dr}(\alpha) = \alpha$ si et seulement si il y a répartition égalitaire.

Donc $m_{dr}(\alpha) \neq \alpha$ (c'est-à-dire $m_{dr}(\alpha) > \alpha$) traduit l'existence d'une répartition non égalitaire donc qu'il y a concentration (voir 2.2.6). Rappelons qu'à ce niveau de l'exposé la notion de concentration est restée qualitative : il y a concentration ou pas. S'il m'est arrivé de parler de moyenne ou forte concentration c'est en faisant référence au coefficient de Gini utilisé par certains pour quantifier justement cette concentration. Mais le but de cet exposé est de montrer que quantifier la concentration d'une série par uniquement g (ou e) n'est pas vraiment une bonne solution.

L'effectif α étant toujours fixé dans $]0;1[$ si $m_{dr}(\alpha)$ augmente cela signifie que la masse possédée par les α derniers individus augmente : ***la masse se concentre (au sens habituel)*** de plus en plus sur les α derniers individus.
J'adopterai donc la définition suivante :

8.1.1 Définition

‖ α étant fixé dans $]0;1[$ on appelera indicateur de concentration du groupe $G_{dr}(\alpha)$
‖ (groupe des α derniers individus)
‖ une quantité c ayant les deux propriétés suivantes :
‖ si s est sa plus petite valeur (sur l'ensemble de toutes les séries) alors
‖ 1) $c = s$ équivaut à répartition égalitaire (c'est à dire $p = 1$)
‖ 2) c varie dans le même sens que $m_{dr}(\alpha)$, c'est à dire c augmente équivaut à
‖ $m_{dr}(\alpha)$ augmente, c'est-à-dire c augmente équivaut à ce que la masse se concentre
‖ de plus en plus (cela au sens habituel) sur les individus du groupe $G_{dr}(\alpha)$.

8.1.2 Propriété

‖ α étant fixé dans $]0;1[$ $m_{dr}(\alpha)$ est un indicateur de concentration du groupe $G_{dr}(\alpha)$

Preuve :

$m_{dr}(\alpha)$ est évidemment une fonction strictement croissante de $m_{dr}(\alpha)$ et sa plus petite valeur est α telle que $m_{dr}(\alpha) = \alpha$, condition équivalente à l'existence d'une répartition égalitaire.

8.1.3 Différentes sortes d'indicateurs de concentration des groupes $G_{dr}(\alpha)$

$m_{dr}(\alpha)$ étant un indicateur de concentration du groupe $G_{dr}(\alpha)$, il suffit de le transformer par une fonction f strictement croissante pour obtenir un autre indicateur de concentration.

***Exemples avec** f **affine et strictement croissante**,* c'est-à-dire $f(x) = ax + b$ et $a > 0$

1) si $a = 1$ et $b = -\alpha$ on obtient $m_{dr}(\alpha) - \alpha$ qui peut s'interpréter comme le surplus de la masse possédée par rapport à une répartition égalitaire ; la plus petite valeur de cet indicateur de concentration du groupe $G_{dr}(\alpha)$ est 0.

2) si $a = \frac{4}{3}$ et $b = \frac{-2}{3}$ on obtient $\frac{4}{3}(m_{dr}(0,5) - 0,5)$, indicateur de concentration du groupe $G_{dr}(0,5)$, c'est-à-dire du groupe des 50% derniers individus. ***En fait le coefficient de Gini est une valeur approchée de cet indicateur. (voir 4.6)***

3) si $a = \frac{1}{\alpha}$ et $b = 0$ on obtient $\frac{m_{dr}(\alpha)}{\alpha} = mse(G_{dr}(\alpha))$ ***qui est donc un indicateur de concentration du groupe*** $G_{dr}(\alpha)$, ***sa plus petite valeur étant*** 1.

***Exemple avec** f **non affine** :*

si $f(x) = \frac{x}{1-x}$ (qui est strictement croissante sur $[0;1[$) on obtient comme indicateur de concentration du groupe $G_{dr}(\alpha)$ l'expression $\frac{m_{dr}(\alpha)}{m_{pr}(1-\alpha)}$ qui n'est autre que le rapport entre la masse possédée par les α derniers individus et la masse possédée par les $1 - \alpha$ premiers individus.

Parmi toutes ces possibilités quel indicateur choisir?

Compte tenu que le passage de l'un à l'autre est immédiat, que les chapitres précédents montrent le rôle important des *mse*, ***lesquels ont par ailleurs une interprétation extrémement concrète (un*** *mse* ***de 2 signifie que le groupe correspondant a 2 fois plus en masse qu'en effectif) je choisirai comme indicateur de concentration des groupes*** $G_{dr}(\alpha)$ ***leurs*** *mse*

8.1.4 Sur les indicateurs de concentration des groupes $G_{pr}(\alpha)$

La masse du groupe $G_{pr}(\alpha)$ est toujours inférieure ou égale à α et elle ne peut prendre la valeur α que s'il y a répartition égalitaire (pour $\alpha \in]0;1[$). Mais c'est lorsque $m_{pr}(\alpha)$ diminue qu'il y a en fait concentration (au sens habituel) de la masse sur le groupe des $1 - \alpha$ derniers individus.

Tout indicateur de concentration des $1 - \alpha$ derniers individus peut donc être considéré comme indicateur de concentration des α premiers individus (s'intéresser aux 60% premiers c'est s'intéresser aux 40% derniers) : ***pour cette raison on ne considérera que les indicateurs de concentration des groupes*** $G_{dr}(\alpha)$.

Une autre raison de s'intéresser surtout aux groupes $G_{dr}(\alpha)$, et c'est peut être la raison la plus importante, est que la propriété 7.5 montre que ce sont ces groupes qui ont les *mse* les plus élevés, ce qui correspond tout à fait à l'objectif initial qui est de rechercher l'existence de groupes ayant beaucoup plus en masse qu'en effectif, c'est -à-dire des groupes ayant justement des *mse* élévés.

8.2 Vecteur concentration

Pour chaque groupe $G_{dr}(\alpha)$ on dispose donc d'un indicateur de concentration : son *mse*. ***L'ensemble de ces*** *mse* ***peut être considéré, par définition, comme une mesure de la concentration de la série.***

Et donc connaître la concentration de cette série, c'est connaître théoriquement tous ces *mse*. Bien entendu en pratique il n'est pas question de déterminer tous les *mse* des groupes $G_{dr}(\alpha)$: il faut se limiter aux *mse* les plus significatifs.

Précisons cet aspect.

Il a été vu au 3.3 qu'une série est caractérisée, à 2 facteurs d'échelle près, par sa courbe de Lorentz ; de façon plus précise cela signifie que la connaissance de la courbe de Lorentz permet de connaître les valeurs du caractère à un coefficient multipicatif prés, et de connaître exactement les fréquences et donc les effectifs à un coefficient multiplicatif près.

Chercher les *mse* les plus significatifs revient donc à chercher des *mse* qui permettent une ‹‹certaine›› localisation de la courbe de Lorentz.

Or justemment, la propriété 4.4 montre que la courbe de Lorentz est relativement localisée lorsqu'on en connaît 2 points autres que $O(0,0)$ et $A(1,1)$. Certes avec 3 points on arriverait à une meilleure localisation, mais le but n'est pas de retouver exactement la série (ce qui serait illusoire) mais d'obtenir une information suffisamment significative en termes de concentration, cela sans trop de calculs.

Essayons donc de choisir deux points dont la connaissance permettrait de localiser suffisamment la courbe C_{pr} de

Lorentz.

Le point d'abcisse centrale $0,5$ paraît incontournable, d'autant plus que bien souvent le coefficient de Gini n'est autre (à une transformation affine près, voir 4.6) que la masse possédée justemment par les 50% premiers individus. L'ordonnée de ce point est $m_{pr}(0,5)$.

Ce point étant choisi, ***la figure de la remarque 2 de 4.4 montre clairement que*** si $m_{pr}(0,5) \simeq 0,5$ alors la courbe de Lorentz est proche de $[OA]$ (ligne d'équirépartition) et il n'y a pas besoin d'un autre point pour localiser C_{pr}.

Si $m_{pr}(0,5) \simeq 0,25$ l'indétermination de C_{pr} est à peu près équivalente à gauche ou à droite de l'abscisse 0,5 (car les triangles OIU et UKA ont des aires sensiblement égales), par contre si $m_{pr}(0,5) \simeq 0$ la courbe de Lorentz est presque entièrement déterminée pour les abscisses inférieures à $0,5$ (puisque C_{pr} est proche de l'axe des abcisses) mais pour les abcisses supérieures à $0,5$ il y a beaucoup d'indétermination.

Il y a donc lieu de choisir comme deuxième point un point d'abscisse supérieure à 0,5. Quelle abscisse choisir?

La loi de Pareto affirme que bien souvent 20% des causes produisent 80% des effets : par exemple 20% des articles vendus contribuent à 80% du chiffre d'affaires, ce qui revient à dire que les 80% premiers articles ont une masse de 20% (du CA) et donc que la courbe de Lorentz passe par le point $(0,8\ ,0,2)$: on pourrait songer à considérer le point d'abcisse 0,8.

Mais en fait la courbe de Lorentz va surtout être déterminée par sa pente finale, c'est-à-dire par la pente de (AM) pour M proche de $A(1,1)$: le point d'abscisse 0,8 me semble trop éloigné pour jouer ce rôle, par contre le point d'abscisse 0,9 me semble plus pertinent. D'autant plus que connaître le point d'abscisse 0,9, c'est connaître la masse des 90% premiers et donc la masse des 10% derniers, quantité souvent citée justemment pour faire remarquer que la masse est concentrée sur un faible effectif.

Je choisirai donc comme deuxième point, le point d'abscisse 0,9.

Notons que les points d'abscisse 0,5 et 0,9 étant choisis, lorsque $m_{pr}(0,5)$ sera voisin de 0,25 la localisation de C_{pr} pour des abscisses inférieures à 0,5 sera évidemment meilleure que dans le cas où on ne connaîtrait que le point d'abscisse 0,5.

On pourrait songer cependant à choisir un 3iéme point, par exemple celui d'abscisse 0,1 (pour raison de symétrie).

Supposons donc connus ces 3 points.

On connaît donc $m_{pr}(0,1),\ m_{pr}(0,5),\ m_{pr}(0,1),$ donc on connaît $m_{dr}(0,9),\ m_{dr}(0,5),\ m_{dr}(0,9)$ ainsi que :

$mse(G_{dr}(0,9)) = \frac{10}{9} m_{dr}(0,9)$

$mse(G_{dr}(0,5)) = 2m_{dr}(0,5)$

$mse(G_{dr}(0,1)) = 10m_{dr}(0,1)$

Mais d'après le point 5 de la propriété 4.4 :

$mse(G_{dr}(0,9))$ décrit tout l'intervalle $\left[1; \frac{10}{9}\right[$ soit une variation possible de 11%

$mse(G_{dr}(0,5))$ décrit tout l'intervalle $[1;2[$ soit une variation possible de 200%

$mse(G_{dr}(0,1))$ décrit tout l'intervalle $[1;10[$ soit une variation possible de 1000%

On peut donc considérer que $mse(G_{dr}(0,9))$ apporte beaucoup moins d'information que les 2 autres indicateurs, cela parce que les 10% premiers ayant toujours peu (au plus 10%), la masse possédée par les 90% derniers est susceptible de peu de variation.

Finalement pour représenter la concentration d'une série on se contentera des deux indicateurs de concentration des groupes $G_{dr}(0,5)$ ***et*** $G_{dr}(0,1)$.

Le couple $(mse(G_{dr}(0,5)), mse(G_{dr}(0,1)))$ ***sans être une caractéristique de la série, est très représentatif de la série. En effet deux séries ayant le même couple*** $(mse(G_{dr}(0,5)), mse(G_{dr}(0,1)))$ ***auront des courbes de Lorentz passant par 4 mêmes points*** :

les points $O(0,0), A(1,1)$ ***et les points*** $(0,5\ ,1 - \frac{mse(G_{dr}(0,5))}{2})$ ***et*** $(0,9\ ,1 - \frac{mse(G_{dr}(0,9))}{10})$.

Donc les courbes de Lorentz de ces deux séries, sans être identiques (il pourra exister des différences notables), ne seront pas fondamentalement différentes : par conséquent il en sera de même pour les fonctions de répartition de la masse m_{pr} ***et*** m_{dr}, ***cela parce que une courbe de Lorentz est caractéristique de la série à deux facteurs d'échelle près.***

Rappelons tout de suite certaines propriétés de ces deux indicateurs en les complétant par d'autres.

8.2.1 Propriété

‖ On notera $c_{50} = mse(G_{dr}(0,5)) = 2m_{dr}(0,5)$ et $c_{10} = mse(G_{dr}(0,1)) = 10m_{dr}(0,1)$

‖ (50 pour 50% derniers et 10 pour 10% derniers)

1) ‖ $1 \leq c_{50} < 2$; $1 \leq c_{10} < 10$

2) ‖ $1 \leq c_{50} \leq c_{10} \leq 9c_{50} - 8 < 10$

‖ $\frac{8 + c_{10}}{9} \leq c_{50} < \min(2, c_{10})$

3) ‖ $c_{50} = 1 \Leftrightarrow c_{10} = 1 \Leftrightarrow$ la répartition est égalitaire

4) ‖ $c_{50} = c_{10} \Leftrightarrow$ il ya répartition égalitaire à l'intérieur des 50% premiers derniers

‖. (les 50% derniers individus ont la même valeur du caractère)

5) ‖ $c_{10} = 9c_{50} - 8 \Leftrightarrow$ il y a répartition égalitaire à l'intérieur des 90% premiers individus.

‖ (les 90% premiers individus ont la même valeur du caractère)

Preuve :

1) Voir le 1 de 7.6 : on fait $\alpha = 0,5$ puis 0,1.

2) L'inégalité $c_{50} \leq c_{10}$ résulte de 7.2

Pour prouver $c_{10} \leq 9c_{50} - 8$. utilisons le 2 de la propriété 4.4 avec $U(0,5, m_{pr}(0,5))$ et $V(0,9, m_{pr}(0,9))$:

$pente[OU] \leq pente[UV] \leq pente[VA]$

$$\frac{m_{pr}(0,5)}{0,5} \leq \frac{m_{pr}(0,9) - m_{pr}(0,5)}{0,4} \leq \frac{1 - m_{pr}(0,9)}{0,1}$$

$$2(1 - m_{dr}(0,5)) \leq \frac{m_{dr}(0,5) - m_{dr}(0,1)}{0,4} \leq 10m_{dr}(0,1)$$

Mais $c_{50} = 2m_{dr}(0,5)$ et $c_{10} = 10m_{dr}(0,1)$ et donc

$$2 - c_{50} \leq \frac{0,5c_{50} - 0,1c_{10}}{0,4} \leq c_{10}$$

La deuxième inégalité redonne $c_{50} \leq c_{10}$, mais la première donne $c_{10} \leq 9c_{50} - 8$.

Quant à l'encadrement de c_{50} en fonction de c_{10}, il est immédiat à obtenir.

3) Il a été vu au 4.3 que pour $\alpha \in]0;1[$, l'égalité $m_{dr}(\alpha) = \alpha$ est caractéristique d'une répartition égalitaire et donc :

$mse(G_{dr}(\alpha)) = 1 \Leftrightarrow$ la répartition est égalitaire

ce qui donne $c_{50} = 1 \Leftrightarrow c_{10} = 1 \Leftrightarrow$ la répartition est égalitaire.

Notons que la triple inégalité $1 \leq c_{50} \leq c_{10} \leq 9c_{50} - 8$ permet de retrouver $c_{50} = 1 \Leftrightarrow c_{10} = 1$

4) $c_{50} = c_{10}$ signifie $mse(G_{pr}(0,5)) = mse(G_{pr}(0,1))$ et par application de 7.3 on obtient le résultat annoncé.

5) D'après la preuve du 2 ci-dessus on a :

$c_{10} = 9c_{50} - 8 \Leftrightarrow pente[OU] = pente[UV] \Leftrightarrow pente[OU] = pente[OV]$ et donc

$c_{10} = 9c_{50} - 8 \Leftrightarrow mse(G_{pr}(0,5)) = mse(G_{pr}(0,9))$ et par application de la propriété 7.3 on obtient le résultat annoncé.

8.2.2 Sur les liens entre c_{50} et c_{10}

La plage de variation de c_{10} à c_{50} fixé est $[c_{50}; 9c_{50} - 8]$ de longueur $8(c_{50} - 1)$

si $c_{50} = 1$, la plage de variation de c_{10} est de longueur nulle puisque $c_{10} = 1$

si $c_{50} = 1,1$ la plage de variation de c_{10} est $[1,1; 1,9]$ de longueur $0,8$

si $c_{50} = 1,5$ la plage de variation de c_{10} est $[1,5; 5,5]$ de longueur 4

si $c_{50} \simeq 2$ la plage de variation de c_{10} est approximativement $[2; 10]$ de longueur 8.

Notons que si $c_{50} = 2$, ce qui signifie que les 50% derniers ont tout (possible que si $x_1 = 0$ et $p \geq 2$), alors c_{10} peut prendre la valeur 10 s'il y a répartition égalitaire au niveau des 50% derniers (voir le 4 de la propriété précédente).

La plage de variation de c_{50} à c_{10} fixé est $\left[\frac{8 + c_{10}}{9}; \min(2, c_{10})\right]$ dont la longueur est $\frac{8(c_{10} - 1)}{9} < 0,9$ si $c_{10} \leq 2$ et de longueur $\frac{10 - c_{10}}{9} < 0,9$ si $c_{10} \geq 2$.

si $c_{10} = 1$ la plage de variation de c_{50} est de longueur nulle puisque $c_{50} = 1$

si $c_{10} = 1,1$ la plage de variation de c_{50} est environ $[1,01;1,1]$ de longueur $0,09$

si $c_{10} = 5,5$ la plage de variation de c_{50} est $[1,5;2]$ de longueur $0,5$

si $c_{10} \simeq 10$ la plage de variation de c_{50} est presque de longueur nulle puisque $c_{50} \simeq 2$

Notons que si $c_{10} = 10$, ce qui signifie que les 10% derniers ont tout (possible si $x_1 = 0$ et $p \geq 2$), alors les 50% derniers ont tout et $c_{50} = 2$.

Analyse des liens entre c_{50} et c_{10}

Dans les 3 cas suivants la connaissance de l'un donne l'autre :

si $c_{10} = 1$ alors $c_{50} = 1$

si $c_{50} = 1$ alors $c_{10} = 1$

si $c_{10} \simeq 10$ alors $c_{50} \simeq 2$.

Mais

1) ***S'il n'y a pas concentration au niveau des 50% derniers individus*** *($c_{50} = 1,1$* ***par exemple) il peut exister cependant une concentration (légère) au niveau des 10% derniers*** puisque c_{10} peut frôler la valeur 2, ce qui signifie que les 10% derniers ont presque 2 fois plus en masse qu'en effectif (cependant 2 reste faible devant la valeur maximum 10 de c_{10}).

2) ***Si la concentration au niveau des 50% derniers est maximale*** *($c_{50} \simeq 2$)* ***alors la concentration au niveau des 10% derniers peut varier beaucoup*** *(c_{10}* ***entre 2 et 10), ce qui recouvre des situations assez différentes***.

3) ***Si l'un a une valeur moyenne, l'autre a une plage de variation réduite de moitié mais qui reste importante*** : si $c_{50} \simeq 1,5$ alors c_{10} peut varier entre 1,5 et 5,5 (il est forcément inférieur à sa valeur moyenne), si $c_{10} \simeq 5,5$ alors c_{50} peut varier entre 1,5 et 2 (il est forcément supérieur à sa valeur moyenne)

(on peut vérifier cet aspect sur les exemples du 6.3 : pour l'exemple 4 on a $c_{50} = 1,44$ et $c_{10} = 2,5$ alors que pour l'exemple 5 on a $c_{50} = 1,52$ et $c_{10} = 1,55$)

Ces 3 aspects montrent clairement que les indicateurs de concentration c_{50} et c_{10} sont relativement indépendants, et ne considérer que c_{50} (par exemple) c'est perdre beaucoup d'information.

8.2.3 conclusion

Si on peut définir une mesure de la concentration d'une série par l'ensemble des *mse* des groupes $G_{dr}(\alpha)$, l'étude faite au 8.2 montre que deux d'entre eux ($mse(G_{dr}(0.5))$ et $mse(G_{dr}(0.1))$) sont particulièrement significatifs :

On peut représenter la concentration d'une série par un vecteur a 2 composantes :
le vecteur concentration $\vec{C}(c_{50}, c_{10})$ avec

$c_{50} = mse(G_{dr}(0.5)) = 2m_{dr}(0.5)$	=indicateur de concentration des 50% derniers individus.
$c_{10} = mse(G_{dr}(0.1)) = 10m_{dr}(0.1)$	=indicateur de concentration des 10% derniers individus

On peut considérer que l'information ‹‹globale›› donnée par la composante c_{50} est complétée par l'information ‹‹finale›› donnée par la composante c_{10}, un peu comme l'écart-type d'une série vient compléter l'information apportée par la valeur moyenne.

On remarquera aussi que dans le cas de séries décilées les deux composantes de ce vecteur concentration sont immédiates à calculer : double de la somme des masses des 5 derniers classes et 10 fois la masse de la dernière classe.

Le chapitre suivant va détailler la mise en oeuvre pratique de cette nouvelle méthode d'analyse de la concentration d'une série à partir du vecteur concentration $\vec{C}$, méthode illustrée sur plusieurs exemples. Au préalable terminons ce chapitre par une comparaison entre le coefficient de Gini g et le vecteur concentration $\vec{C}$.

8.3 Lien entre le coefficient de Gini g et le vecteur concentration $\vec{C}$

Il a été vu au 4.4 que $g = 2\int_0^1 (m_{dr}(\alpha) - \alpha)d\alpha = 2\int_0^1 m_{dr}(\alpha)d\alpha - 1$, c'est-à-dire g est le double de la valeur moyenne de tous les indicateurs de concentration $m_{dr}(\alpha) - \alpha$.

Ceci explique déjà pourquoi la connaissance de g ne permet de rien dire de précis sur la série, du moins en terme de *mse* (tout comme la valeur moyenne d'une série ne permet pas de connaître les nombres de la série).

Allons plus loin dans l'analyse.

D'après le 4.6 $\boxed{g \simeq \frac{4}{3}(m_{dr}(0,5) - 0,5) = \frac{2}{3}(c_{50} - 1)}$

Le coefficient de Gini est donc pratiquement égal, à une transformation affine près, à la première composante c_{50} du vecteur concentration $\vec{C}$: il traduit donc surtout ce qui se passe au niveau des 50% derniers. Cela explique en particulier pourquoi g peut être trés faible (0,075) alors qu'un groupe (les 9% derniers) a 1,75 fois plus en masse qu'en effectif, ce qui n'est tout de même pas une répartition presque égalitaire (voir exemple du 3.2.2).

Finalement, le coefficient de Gini au prix de calcul compliqués remplace le vecteur concentration $\vec{C}$ par une seule de ses deux composantes ce qui évidemment conduit à une perte d'information!

En conclusion :

le vecteur concentration $\vec{C}$, avec beaucoup moins de calculs,
apporte beaucoup plus d'informations que le coefficient g
et donc va permettre une analyse beaucoup plus fine de la série en terme de concentration.

Remarque 1 :

Si on pose $c = \frac{c_{50} + c_{10} - 2}{10}$ alors

1) $c \in [0;1]$

2) $c = 0 \Leftrightarrow c_{50} = 1$ et $c_{10} = 1 \Leftrightarrow$ la répartition est égalitaire.

3) $c = 1 \Leftrightarrow c_{50} = 2$ et $c_{10} = 10 \Leftrightarrow$ les 10% derniers ont tout (si $x_1 = 0$ et $p \geq 2$)

Les deux premiers points résultent de la propriété 8.2.1. Le troisiéme résulte du point 4 de 7.6.

Ce coefficient c a donc pratiquement le même comportement que le coefficient de Gini : il varie entre 0 et 1, la valeur 0 caractérisant la répartition égalitaire et la valeur 1 caractérisant une concentration maximum au niveau des 10% derniers (qui ont alors tout, ainsi que les 50% derniers).

Mais ne considérer que c, c'est perdre de l'information par rapport à $\vec{C}$

Remarque 2 :

On pourrait envisager de calculer la valeur moyenne de tous les *mse*, c'est à dire l'intégrale $\int_0^1 \frac{m_{dr}(\alpha)}{\alpha}d\alpha$, mais ce ne serait là aussi qu'une valeur moyenne donc cela conduirait encore à une perte d'information. En outre la complexité du résultat (voir annexe 6) ôte toute envie de l'utiliser en pratique.

9
Méthode *mse* d'analyse de la concentration d'une série et applications à de nombreux exemple

Dans le 9.1 et 9.2 la courbe de Lorentz (définie au 3.1.1 et 3.1.2) sera évoquée 3 fois : on peut dans un premier temps ignorer les passages correspondants car ils ne sont pas nécessaires à la mise en oeuvre pratique de la méthode *mse*, ils en apportent simplement un éclairage géométrique.

9.1 Rappel de notations et de résultats.

Il s'agit ici de résumer les notations utilisées dans les chapitres précédents et les principaux résultats correspondants, cela surtout pour les lecteurs qui sont passés directement de l'introduction à ce chapitre.

On considére une série d'individus dont on étudie un caractère quantitatif prenant les valeurs $x_1, x_2, \ldots, x_p$ avec $p \geq 1$, $x_1 > 0$, $x_i < x_{i+1}$. Dans un cas continu les milieux des classes seront les x_i et on supposera que tous les individus d'une classe ont comme valeur du caractère le milieu de la classe. Les effectifs correspondants à ces valeurs du caractère seront notés $n_1, n_2, \ldots, n_p$ ***tous strictement positifs***.

On posera :

$\boxed{n = \sum_{i=1}^{p} n_i}$ l'effectif total.

$\boxed{m = \sum_{i=1}^{p} n_i x_i}$ la masse totale (par exemple la masse salariale d'une entreprise dans le cas où les individus sont les employés de cette entreprise, le caractère étant le salaire de l'employé).

$\boxed{\bar{x} = \frac{m}{n}}$ désignera la moyenne de la série.($\bar{x} > 0$).

On pourra parfois envisager le cas $x_1 = 0$ mais pour $p \geq 2$ (ce qui assure encore que $m > 0$ et $\bar{x} > 0$), mais cela sera alors explicitement dit.

On cherche à savoir s'il existe un (ou plusieurs) groupe(s) d'individus possédant une part de la masse nettement supérieure à son poids démographique, c'est-à-dire si la masse est ***concentrée*** sur un (ou plusieurs) groupe(s) particulier(s).

Afin de pouvoir illustrer immédiatement les différents résultats qui vont être énoncés considérons l'exemple d'une série de 80 individus (des terres agricoles), le caractère étant la superficie en hectares de chaque terre agricole :

x_i	n_i	$n_i x_i$	$\Sigma n_i \searrow$	$\Sigma n_i x_i \searrow$
5	16	80	80	2130
15	30	450	64	2050
30	18	540	34	1600
55	10	550	16	1060
85	6	510	6	510
	$n = 80$	$m = 2130$		

Les dernières colonnes sont constituées des effectifs et masses en cumulés décroissants : leur utilité apparaîtra

ci-aprés. Elles s'interprètent ainsi : les 6 derniers individus possèdent 510 hectares, les 16 derniers individus possèdent 1060 hectares.....
La valeur moyenne de cette série, c'est à dire la valeur moyenne des superficies des 80 terres agricoles est $\bar{x} = \frac{2130}{80} = 26,625$ hectares.

1) Notion de *mse*

Pour tout groupe d'individus G d'effectif non nul on définit son rapport masse sur effectif (*mse*) par :

$$mse(G) = \frac{\text{masse (en pourcentage de la masse totale) possédée par le groupe } G}{\text{effectif (en pourcentage de l'effectif total) du groupe } G}$$

On en déduit $$mse(G) = \frac{\text{moyenne du caractère pour le groupe } G}{\text{moyenne du caractère pour toute la série}} = \frac{\bar{x}_G}{\bar{x}}$$ (voir 2.2.2)

et donc on a toujours $mse(G)$ compris entre $\frac{x_1}{\bar{x}}$ et $\frac{x_p}{\bar{x}}$

Le plus grand *mse* est évidemment réalisé par tout sous-groupe des n_p derniers individus .

(voir 2.2.3 pour ces 2 derniers résultats)

Ainsi, dans l'exemple considéré le *mse* des 16 derniers individus (20% de l'effectif total) est

$$\frac{\frac{550+510}{2130}}{\frac{16}{80}} = \frac{\frac{550+510}{16}}{\frac{2130}{80}} \simeq 2,49$$

c'est à dire ce groupe des 16 derniers individus a pratiquement 2,5 fois plus en masse qu'en effectif.
Cette valeur est bien comprise entre le plus petit *mse* ($\frac{x_1}{\bar{x}} = \frac{5}{26,625} \simeq 0,19$) et le plus grand *mse* ($\frac{x_5}{\bar{x}} = \frac{85}{26,625} \simeq 3,19$), ce dernier étant le *mse* des 6 derniers individus (7,5% de l'effectif total).

.
La considération de ces** mse **est justifiée par le fait qu'il est tout de même ***plus parlant*** de dire qu'un groupe possède 2,5 fois plus en masse qu'en effectif que de dire qu'il possède une masse supérieure de 0,3 à son effectif (50%-20%, puisqu'ici le groupe des 16 derniers individus, soit 20% de l'effectif total possède $\frac{1060}{2130} \simeq 50\%$ de la masse totale).

2)Répartition égalitaire et *mse*

Une répartition est égalitaire signifie que tout groupe d'individus a ***autant en masse qu'en effectif***, c'est-à-dire que tout groupe a un *mse* est égal à 1 : dans ce cas ***il n'y a pas de phénomène de concentration***.
Cette situation a lieu (voir 2.2.5) si et seulement si ***tous les individus ont la même valeur du caractère*** ($p = 1$).

3) Sur le *mse* des groupes constitués des derniers individus

Les individus de la série étant classés par valeur croissante du caractère et α étant dans $[0;1]$ (α représentera toujours dans ce qui suit un effectif exprimé en pourcentage de l'effectif total), on définit :

$G_{dr}(\alpha) =$ groupe constitué des α derniers individus

$m_{dr}(\alpha) =$ masse (en pourcentage) possédée par le groupe des α derniers individus

et donc si $\alpha \in]0;1]$ on a

$$\boxed{mse(G_{dr}(\alpha)) \ = \frac{m_{dr}(\alpha)}{\alpha}}$$

.

La nécessité de considérer ces groupes est justifiée par le point suivant : l'objectif est de chercher l'existence de groupes ayant beaucoup plus en masse qu'en effectif, donc ayant des mse élevés, or les groupes ayant les mse les plus élevés sont justement les groupes constitués de derniers individus d'aprés 7.5!

Toujours pour l'exemple considéré on a :

$$mse(G_{dr}(0,5)) \ = \frac{\frac{1600 + (40-34)\times 15}{2130}}{0,5} \simeq 1,58$$

(puisque 50% de l'effectif total est 40 et on exploite les colonnes des effectifs et masse en cumulés décroissants)

$$mse(G_{dr}(0,1)) \ = \frac{\frac{510 + (8-6)\times 55}{2130}}{0.1} \simeq 2,9$$

(puisque 10% de l'effectif total est 8)

Remarque :

La valeur maximale de $\frac{m_{dr}(\alpha)}{\alpha}$ lorsque α décrit $]0;1]$ est $\frac{x_p}{\bar{x}}$ (puisque le plus grand *mse* est $\frac{x_p}{\bar{x}}$, valeur réalisée par le groupe des n_p derniers individus) ;

La valeur maximale de $m_{dr}(\alpha) - \alpha$ a été étudiée au chapitre 5.

4) *mse* et concentration

Il a été vu au 8.1 que pour tout $\alpha \in]0;1[$ $mse(G_{dr}(\alpha))$ ***est un indicateur de concentration du groupe des α derniers individus*** : en effet α restant fixe, ***si ce mse augmente*** c'est bien que la masse possédée par ce groupe augmente et donc que ***la masse se concentre (au sens habituel) de plus en plus sur ce groupe***.

Par ailleurs le 4.3 montre, (puisque $mse(G_{dr}(\alpha)) \ = \frac{m_{dr}(\alpha)}{\alpha}$), que pour α quelconque dans $]0;1[$ on a :

$$\boxed{mse(G_{dr}(\alpha)) \ \geq 1 \text{ et } mse(G_{dr}(\alpha)) \ = 1 \Leftrightarrow \text{la répartition est égalitaire}}$$

c'est-à-dire

UN seul de ces *mse* permet de caractériser le fait qu'il y ait répartition égalitaire ou pas

On peut considérer que connaître l'état de concentration de la série, c'est connaître à priori tous ces *mse*. Mais évidemment il n'est pas question de les calculer tous : en fait le 8.2 montre que l'on peut se contenter de déterminer le *mse* des 50% derniers et le *mse* des 10% derniers. En effet toutes les séries ayant respectivement ces deux mêmes *mse* auront des courbes de Lorentz passant par 4 mêmes points, et donc sans être identiques (il pourra exister des différences notables) ces courbes ne seront pas fondamentalement différentes : par conséquent il en sera de même pour la fonction de répartition de la masse m_{dr}.

On représentera donc la concentration d'une série par le vecteur $\overrightarrow{C}$ dont les deux composantes sont justement ces deux *mse* : on peut considérer que ***l'information ‹‹globale››*** donnée par le *mse* des 50% derniers est complétée par ***l'information ‹‹finale››*** donnée par le *mse* des 10% derniers, un peu comme l'écart-type d'une série vient compléter l'information apportée par la valeur moyenne.

9.2 Méthode *MSE* d'analyse de la concentration d'une série.

Tout d'abord si $\frac{x_p}{x_1} \leq 1,25$ il est inutile d'aller plus loin : la répartition est presque égalitaire, il n'y a pas de

phénomène de concentration, puisque tout groupe d'effectif supérieur à 50% a un *mse* inférieur ou égal à 1,11 et tout groupe d'effectif supérieur ou égal à 10% a un *mse* inférieur ou égal à 1,22 (voir remarque 4 du 9.2.2 ci-aprés).

Dans le cas contraire, il y a lieu évidemment de poursuivre l'analyse :

9.2.1 Phase 1 :calcul du vecteur concentration $\vec{C}(c_{50}, c_{10})$

$c_{50} = mse(G_{dr}(0,5)) = 2m_{dr}(0,5)$ = indicateur de concentration des 50% derniers individus.
$c_{10} = mse(G_{dr}(0,1)) = 10m_{dr}(0,1)$ = indicateur de concentration des 10% derniers individus

Rappel : $m_{dr}(\alpha)$ est la masse (en pourcentage de la masse totale) possédée par le groupe des α (en pourcentage de l'effectif total) derniers individus.

On qualifiera c_{50} d'indicateur de ***concentration globale*** et c_{10} d'indicateur de ***concentration finale***.

On a toujours (il s'agit d'encadrements valables pour TOUTES les séries)

$$1 \leq c_{50} < 2 \qquad 1 \leq c_{10} < 10$$

$$1 \leq c_{50} \leq c_{10} \leq 9c_{50} - 8 < 10 \quad \text{(d'aprés 8.2.1)}$$

On notera que dans le cas de séries décilées c_{50} n'est autre que le double de la somme des masses (en pourcentage) des 5 dernières classes et c_{10} est 10 fois la masse (en pourcentage) de la dernière classe. Les calculs sont donc on ne peut plus simples, il n'y a même pas à faire les 2 colonnes des effectifs et masse en cumulés décroissants.

Dans le cas de séries non décilées la détermination des composantes de $\vec{C}$ reste légère : les calculs ci-dessus faits pour l'exemple des terres agricoles (aucune étape escamotée) le montrent clairement. On trouve $c_{50} = 1,58$ et $c_{10} = 2,9$.

Si 50%n et 10%n ne sont pas des entiers le principe de calcul reste le même : par exemple si on avait obtenu $50\%n = 40,4$ on aurait considéré (40,4-34)×15, ceci correspondant au fait que la fonction m_d est une fonction affine par intervalles (voir 4.2).

Remarque :

On verra à l'exemple 9 du 9.3 une façon de présenter les calculs en utilisant les effectifs et les masses en cumulés décroissants et en pourcentage : le seul interêt de cette façon est de permettre la représentation graphique C_{dr} de la fonction m_{dr}, mais ***cette représentation graphique*** n'apporte, quantitativement, aucun plus par rapport au vecteur concentration $\vec{C}$: ***elle n'est absolument pas indispensable à faire***.

Notons que

a) C_{dr} passe (par définition) par les points $N_{0,5}\ (0,5\ , \frac{c_{50}}{2})$ et $N_{0,1}\ (0,1\ , \frac{c_{10}}{10})$, ce qui permet de donner une interprétation géométrique de c_{50} et c_{10} : ces deux *mse* ne sont autres que les pentes des droites $(ON_{0,5})$ et $(ON_{0,1})$.
Et puisque $m_{dr}(0) = 0$ et $m_{dr}(1) = 1$ il est clair que toutes les séries ayant le même vecteur concentration $\vec{C}$ ont des courbes C_{dr} passant par les 4 mêmes points suivants :
$O(0,0),\ N_{0,1},\ N_{0,5},\ A(1,1)$

b) C_{dr} est la symétrique de la courbe de Lorentz (C_{pr}) par rapport au point $(\frac{1}{2}, \frac{1}{2})$, cette courbe de Lorentz étant la représentation graphique de la fonction $m_{pr}(\alpha) = 1 - m_{dr}(1-\alpha)$ = masse possédée par les α premiers individus.(voir 3.1.2, 4.1, 4.2)

c) C_{pr} passe par les points $M_{0,5}\ (0,5\ ,1 - \frac{c_{50}}{2})$ et $M_{0,9}\ (0,9\ , 1 - \frac{c_{10}}{10})$, et donc toutes les séries ayant le même vecteur concentration $\vec{C}$ ont des courbes de Lorentz passant par les 4 mêmes points $O(0,0),\ M_{0,5},\ M_{0,9},\ A(1,1)$, aspect déjà signalé au 4 du 9.1

9.2.2 Phase 2 :analyse des résultats

Cette analyse repose sur les diverses propriétés de c_{50} et c_{10} : on peut les décomposer en 5 points.

1)

c_{50} est le plus grand *mse* parmi tous les groupes d'effectifs supérieur ou égal à 50% (voir 7.4)

c_{10} est le plus grand *mse* parmi tous les groupes d'effectifs supérieur ou égal à 10% (voir 7.4)

Si la série est décilée ou si l'effectif des n_p derniers individus est supérieur ou égal à 10% alors c_{10} est le plus grand mse pour tous les groupes de la série (voir remarque 1 ci-aprés)

2)

La répartition est égalitaire si et seulement si $c_{50} = 1$ ou si et seulement si $c_{10} = 1$ (voir 8.2.1).

Rappelons que la répartition est égalitaire signifie que tous les individus de la série ont la même valeur du caractère ($p = 1$)et notons que :

a) la répartition est égalitaire équivaut à ce que la représentation graphique C_{dr} de m_{dr} soit le segment $[OA]$, avec $O(0,0)$ et $A(1,1)$, ce qui équivaut à dire que cette représentation graphique a un point commun avec $]OA[$ (d'aprés 4.1, 4.2, 4.3). On a le même résultat avec la courbe de Lorentz.

b) si c_{10} est proche de 1, c_{50} est au moins aussi proche de 1 puisque $1 \leq c_{50} \leq c_{10}$.
si c_{50} est proche de 1, c_{10} peut être un peu plus éloigné de 1 puisque $c_{10} - 1 \leq 9(c_{50} - 1)$, voir exemple 9.3.6.

3)

$c_{50} = c_{10} \Leftrightarrow$ il y a répartition égalitaire à l'intérieur du groupe des 50% derniers

$c_{10} = 9c_{50} - 8 \Leftrightarrow$ il y a répartition égalitaire à l'intérieur du groupe des 90% premiers individus

Voir 8.2.1.

4) L'hypothèse générale $x_1 > 0$ entraîne que c_{10} et c_{50} ne peuvent atteindre les valeurs 2 et 10 (voir 7.6), mais si on accepte la possibilité $x_1 = 0$ (avec $p \geq 2$) alors il est clair que :

$c_{50} = 2$ équivaut à ce que les 50% premiers individus ont 0 ($\alpha = 0,5$ dans le 3 de 7.6)

(car pour que les 50% derniers aient tout, les 50% premiers ne doivent rien avoir, donc ils doivent avoir tous $x_1 = 0$ comme valeur du caractère)

$c_{10} = 10$ équivaut à ce que les 90% premiers individus ont 0 ($\alpha = 0,1$ dans le 3 de 7.6)

(même explication)

On notera que $c_{10} = 10$ entraîne que $c_{50} = 2$, l'inverse étant évidemment faux : si $c_{50} = 2$ (donc $x_1 = 0$) alors $2 \leq c_{10} \leq 10$ et c_{10} ne prendra la valeur 2 que s'il y a répartition égalitaire au niveau des 50% premiers (voir 3 ci-dessus).

On peut faire aussi cette remarque : si $c_{10} \simeq 10$, alors $g \gtrapprox 0,9$. Pour le justifier on peut utiliser l'annexe 8. En effet, puisque $c_{10} \simeq 10$, $c_{50} \simeq 2$, donc $a = 1 - 0,5c_{50} \simeq 0$ et $b = 1 - 0,1c_{10} \simeq 0$ et le minorant $1 - (0,9a + 0,5b + 0,1)$ de g trouvé à l'anexe 8 est $\simeq 0,9$.

5) Compte tenu que les plages de variation (à $x_1 > 0$) de c_{50} et c_{10} sont effectivement $[1;2[$ et $[1;10[$ (voir 7.6), on peut donner des qualificatifs aux diverses situations, par exemple :

Qualificatifs de concentration globale

$c_{50} = 1$	répartition égalitaire ou concentration nulle
$c_{50} \simeq 1.5$	concentration globale moyenne
$c_{50} \simeq 2$	concentration globale maximum : les 50% derniers ont pratiquement tout

Qualificatifs de concentration finale

$c_{10} = 1$	répartition égalitaire ou concentration nulle
$c_{10} \simeq 5,5$	concentration finale moyenne
$c_{10} \simeq 10$	concentration finale maximum : les 10% derniers ont pratiquement tout

Exemple** : **analysons tout de suite les résultats de l'exemple du 9.1 sur les terres agricoles pour lequel en 2 lignes de calculs on a trouvé :

$$c_{50} = 1,58 \text{ et } c_{10} = 2,9$$

Cela signifie que les 50% derniers individus ont presque 1,6 fois plus en masse qu'en effectif alors que les 10% derniers ont eux presque 3 fois plus en masse qu'en effectif. Ces deux informations trés précises peuvent se traduire par :

la ***concentration globale*** est donc un petit peu en dessus de la moyenne, mais la ***concentration finale*** est nettement en dessous de la moyenne.

Avant de passer à d'autres exemples donnons quelques remarques complémentaires qui peuvent être sautées en première lecture :

Remarque 1 :

Comme il l'a été rappelé au 9.1 le plus fort *mse* ($\frac{x_p}{\bar{x}}$) est réalisé par tout sous-groupe des n_p derniers individus, en particulier par le groupe des n_p derniers individus.

***Dans le cas d'une série décilée**,* les n_p derniers individus sont les 10% derniers individus, donc le *mse* des 10% derniers individus (c_{10}) est le plus fort *mse*.

Dans le cas d'une série non décilée mais avec l'effectif des** n_p **derniers individus supérieur ou égal à 10% (ce qui est loin d'être toujours réalisé), alors le groupe des 10% derniers individus est un sous groupe des n_p derniers individus, donc son *mse* (c_{10}) est aussi le plus grand *mse*.

***Dans le cas contraire**, c_{10}* n'est pas le plus grand *mse* de la série : c'est justemment le cas de la série des terres agricoles où le dernier effectif est 7,5%, inférieur à 10% et donc le plus fort *mse* est celui des 7,5% derniers individus soit $3,19$ alors que $c_{10} = 2,9$. Mais pour apprécier ce plus fort *mse* il faut considérer sa plage de variation possible à savoir $\left[1; \frac{1}{7,5\%}\right[$ (voir 7.6) soit $[1;13,3[$ de valeur centrale 7,15, nettement supérieur à 3,19. Ainsi au niveau des 7,5% derniers individus la concentration est nettement en dessous la valeur moyenne .

L'inconvénient du recours à ce plus fort *mse* est qu'il ne permet pas de comparaison facile entre deux séries puisque le pourcentage $\frac{n_p}{n}$ varie d'une série à une autre : c'est une information qui ne me parait pas indispensable de mettre en évidence systématiquement.

Remarque 2 :

Il s'agit ici de préciser les majorations des deux composantes du vecteur concentration $\vec{C}$.

Les majorants de c_{50} et c_{10} donnés ci dessus (2 et 10) sont valables sur l'ensemble de toutes les séries. En fait on peut donner des majorants tenant compte de la série particulière considérée :

c_{50} et c_{10} sont inférieurs ou égaux à $\frac{x_p}{\bar{x}}$ (puisque ce sont des *mse*)

Ce majorant $\frac{x_p}{\bar{x}}$ dépend des effectifs et des valeurs du caractère.

Mais on a aussi, en faisant $\alpha = 0,5$ et $\alpha = 0,1$ dans le 2 et 4 de 7.6 :

$c_{50} \leq \frac{2x_p}{x_1 + x_p}$, majorant qui ne dépend pas des effectifs

L'égalité a lieu si et seulement si $p = 2$ et si les 50% premiers ont x_1 comme valeur du caractère, les autres ayant x_2 comme valeur du caractère.

$c_{10} \leq \frac{10x_p}{9x_1 + x_p}$, majorant qui ne dépend pas des effectifs

L'égalité a lieu si et seulement si $p = 2$ et si les 90% premiers ont x_1 comme valeur du caractère, les autres ayant x_2 comme valeur du caractère.

Remarque 3 :
De $c_{50} \leq \frac{2x_p}{x_1 + x_p}$ et $c_{10} \leq \frac{10x_p}{9x_1 + x_p}$ on obtient (en posant $q = \frac{x_p}{x_1}$) : $c_{50} \leq \frac{2q}{1+q}$ et $c_{10} \leq \frac{10q}{9+q}$, expressions qui sont croissantes avec q :

q	$\frac{2q}{1+q}$	$\frac{10q}{9+q}$
2	1,33	1,82
1,5	1,2	1,43
1,25	1,11	1,22

Donc si $\frac{x_p}{x_1} \leq 1,25$ on est sûr que $c_{50} \leq 1,11$, ce qui assure que d'un point de vue global il n'y a pratiquement pas concentration (puisque tout groupe d'effectif supérieur ou égal à 50% a un $mse \leq 1,11$ donc pratiquement égal à 1). On est également sûr que $c_{10} \leq 1,22$, donc tout groupe d'effectif supérieur ou égal à 10% a un $mse \leq 1,22$. Et même si 1,22 est plus éloigné de 1 que 1,11, on peut considérer qu'il n'y a pratiquement pas concentration d'un point de vue final puisque la valeur maximum de c_{10} est 10.

9.3 Exemples d'analyse en utilisant le vecteur concentration $\vec{C}(c_{50},c_{10})$: méthode *MSE*.

Dans la plupart des exemples, certaines comparaisons seront faites entre cette nouvelle méthode *mse* et la méthode Lorentz-Gini (détaillée au chapitre 3) : elles peuvent évidemment être ignorées par le lecteur ne connaissant pas cette méthode.

9.3.1 Exemple 1

Une entreprise fait un chiffre d'affaires de 2563500KF répartis selon 176 factures :

classes	x_i(milieu)	n_i	$n_i x_i$	$\Sigma n_i \searrow$	$\Sigma n_i x_i \searrow$
$[0;15000[$	7500	139	1042500	176	2563500
$[15000;30000[$	22500	12	270000	37	1521000
$[30000;45000[$	37500	16	600000	25	1251000
$[45000;60000[$	52000	3	156000	9	651000
$[60000;75000[$	67500	2	135000	6	495000
$[75000;90000[$	82500	2	165000	4	360000
$[90000;105000[$	97500	2	195000	2	195000
		$n = 176$	$m = 2563500$		

Déterminons le vecteur concentration $\vec{C}$:
$50\%n = 88$ et puisque $37 < 88 < 176$

$$c_{50} = mse(G_{dr}(0,5)) = 2m_{dr}(0,5) = 2 \times \frac{1521000 + (88 - 37) \times 7500}{2563500} \simeq 1,58$$

$10\%n = 17,6$ et puisque $9 < 17,6 < 25$

$$c_{10} = mse(G_{dr}(0,1)) = 10m_{dr}(0,1) = 10 \times \frac{651000 + (17,6 - 9) \times 37500}{2563500} \simeq 3,8$$

Conclusion :
la concentration globale est un peu au dessus de la moyenne (la valeur moyenne est 1,5) mais ***la concentration finale est en dessous de la moyenne*** (la valeur moyenne est 5,5)

Remarque 1 :

Cette remarque est en liaison avec la remarque 1 du 9.2.2.

Le *mse* le plus élevé n'est pas 3,8 puisque la série n'est pas décilée et l'effectif des $n_p = 2$ derniers individus ($\frac{2}{176} \simeq 1,1\%$ de l'effectif) n'est pas supérieur à 10% : le *mse* le plus élevé est justement le *mse* des 2 derniers individus égal à $\frac{x_p}{\bar{x}} = \frac{97500}{\frac{2563500}{176}} \simeq 6,7.$ Mais la plage de variation (sur l'ensemble de toutes les séries) du *mse* des 1,1% derniers individus est $\left[1; \frac{1}{1,1\%}\right[$, soit environ $[1;91[$, et donc en fait la concentration au niveau du groupe ayant le plus fort *mse* est ici trés faible.

Remarque 2 :

Par rapport à l'exemple du 9.1 (série des terres agricoles) et analysé au 9.2.2, on notera que les c_{50} sont égaux (à 1,58) alors que les c_{10} sont différents (2,9 et 3,8) : la série des factures ci-dessus est un peu plus concentrée au niveau des 10% derniers que celle des terres agricoles, cela pour une même concentration au niveau des 50% derniers.

Remarque 3 :

Je laisse au lecteur le soin de vérifier que le coefficient de Gini de cette série est $g \simeq 0,41$: après beaucoup plus de calculs on ne peut que conclure à ‹‹concentration›› un peu en dessous de la moyenne sans rien savoir de précis sur les groupes ayant plus en masse qu'en effectif.

On peut aussi vérifier le 8.3, à savoir qu'une bonne valeur approchée de g est $\frac{2}{3}(c_{50} - 1) = \frac{2}{3}0,58 \simeq 0,39.$; cette proximité entre $\frac{2}{3}(c_{50} - 1)$ et g sera vérifiée sur d'autres exemples (soyons honnête : il peut arriver que cette approximation ne soit pas aussi bonne).

Ceci montre que calculer le coefficient de Gini,
c'est pratiquement calculer uniquement
la masse possédée par les 50% derniers individus,
cela par un chemin détourné et lourd.

9.3.2 Exemple 2

Source : référence [6] page 149.

x_i	n_i	$n_i x_i$	$\Sigma n_i \searrow$	$\Sigma n_i x_i \searrow$
5	5	25	43	403
7	8	56	38	378
9	12	108	30	322
11	10	110	18	214
13	8	104	8	104
	$n = 43$	$m = 403$		

Déterminons le vecteur concentration $\vec{C}(c_{50}, c_{10})$

$50\%n = 21,5$ et comme $18 < 21,5 < 30$

$c_{50} = mse(G_{dr}(0,5)) = 2m_{dr}(0,5) = 2 \times \frac{204 + (21,5 - 18) \times 9}{403} = 1,2$

$10\%n = 4,3 < 8$ d'où

$c_{10} = mse(G_{dr}(0,1)) = 10m_{dr}(0,1) = 10 \times \frac{13 \times 4,3}{403} \simeq 1,39$

la concentration globale est donc faible (1,2 bien inférieur à la valeur moyenne 1,5) et ***la concentration finale très faible*** (1,39 très inférieur à la valeur moyenne 5,5).

Remarque 1 :
Notons qu'ici 1,39 est le plus fort *mse* de la série puisque l'effectif des n_p derniers individus est supérieur à 10% et donc tout groupe a une masse inférieure ou égale à 1,39 fois son effectif.

On peut vérifier que le *mse* des n_p derniers individus est effectivement c_{10} puisque ce *mse* est égal à $\frac{\frac{104}{403}}{\frac{8}{43}}$.

Remarque 2 :
Je laisse à nouveau au lecteur le soin de faire la courbe de Lorentz et de calculer le coefficient de Gini ($\simeq 0,15$) : au prix de calculs beaucoup plus long il arrivera à une conclusion peu précise, à savoir faible ‹‹concentration››, cela parce que le coefficient de Gini est proche de zéro (ce qui traduit la proximité de la courbe de Lorentz au segment $[OA]$).
Remarque 3 : on est encore ici dans un cas où $\frac{2}{3}(c_{50} - 1) \simeq 0,13$ est une bonne approximation de g (voir 8.3)

9.3.3 Exemple 3 et 4

Source : référence [1] pages 107 et 109, il s'agit des séries des Revenus et Patrimoines 1986 présentées sous forme décilées.

classe	part du revenu (en %)	part du patrimoine (en %)
1	2,2	0,1
2	3,8	0,3
3	4,9	0,8
4	6	1,6
5	7,2	3,2
6	8,8	5,9
7	10,6	8,6
8	12,6	10,6
9	16,1	15,1
10	27,8	53,8

De façon immédiate à partir du tableau ci-dessus on obtient :

pour les revenus

$c_{50} = 2m_{dr}(0,5) = 2 \times \frac{27,8 + 16,1 + 12,6 + 10,6 + 8,8}{100} \simeq 1,52$

$c_{10} = 10m_{dr}(0,1) = 2,78$

la concentration globale est moyenne, la concentration finale est en dessous la moyenne.

pour les patrimoines

$c_{50} = 2m_{dr}(0,5) = 2 \times \frac{53,8 + 15,1 + 10,6 + 8,6 + 5,9}{100} \simeq 1,88$

$c_{10} = 10m_{dr}(0,1) = 5,38$

la concentration globale est extrême (on n'est pas loin de la valeur maximun) et la concentration finale est moyenne.

La série des patrimoines est donc beaucoup plus concentrée que la série des revenus, cela aussi bien au niveau global qu'au niveau final.

Remarque 1 :
Les séries étant décilées c_{10} est le plus fort *mse*.
Par exemple pour la série des revenus tout groupe a une masse inférieure ou égale à 2,78 fois son effectif.

Remarque 2 :

Avec des calculs plus lourds (voir 6.2) la méthode Gini conduit à une conclusion moins nuancée : la série des patrimoines est beaucoup plus ‹‹concentrée›› que la série des revenus puisque $g_{revenus} \simeq 0,37$ et $g_{patrimoines} \simeq 0,66$, mais hélas les chiffres 0,37 et 0,66 n'ont pas vraiment d'interprétation économique précise.

Il est tout même plus parlant de dire, par exemple pour les patrimoines, que les 50% derniers individus ont presque 1,9 fois plus en masse qu'en effectif et que les 10% derniers ont 5,3 fois plus en masse qu'en effectif que de dire $g \simeq 0,66$ (ce qui n'autorise en outre qu'une conclusion moins précise).

Remarque 3 :

Il existe des séries plus concentrées que la série des patrimoines au niveau des 10% derniers, par exemple en 1999 ‹‹65% de l'encours global des actions, obligations et autres OPCVM se trouve logé dans quelques 11% seulement des comptes titres ouverts dans les banques et sociétés de Bourse françaises›› (d'aprés un journal économique de août 1999), donc $mse(G_{dr}(11\%)) = \frac{65}{11} \simeq 5,91$ et d'aprés 7.3 on a $\frac{65}{11} \leq c_{10}$ (il y aura égalité si et seulement si la répartition est égalitaire à l'intérieur du groupe des 11% derniers). Enfin puisque $c_{10} \leq 9c_{50} - 8$ on est sûr que $1,55 \leq c_{50} < 2$.

9.3.4 Exemple 5

Il s'agit encore de la série des revenus mais cette fois de 1994 et présentée sous une forme non décilée:

x_i est le revenu net annuel exprimé en milliers de francs

n_i est l'effectif correspondant exprimé en millions de personnes

$n_i x_i$ est donc exprimé en milliards de francs.

x_i	n_i	$n_i x_i$	$\Sigma n_i \searrow$	$\Sigma n_i x_i \searrow$
30	2,7	81	37,2	4379
42	1	42	34,5	4298
50,39	6,35	320	33,5	4256
84	12,85	1079	27,15	3936
129	8,35	1077	14,3	2857
216	4,2	907	5,95	1780
498,5	1,75	873	1,75	873
	$n = 37,2$	$m = 4379$		

Déterminons le vecteur concentration $\vec{C}(c_{50}, c_{10})$

$50\%n = 18,6$ d'où $c_{50} = 2 \times \frac{2857 + (18,6 - 14,3) \times 84}{4379} \simeq 1,47$

$10\%n = 3,72$ d'où $c_{10} = 10 \times \frac{873 + (3,72 - 1,75) \times 216}{4379} \simeq 2,96$

Donc pour les revenus, de 1986 à 1994, ***la concentration globale est restée à peu près la même*** (c_{50} passe de 1,52 à 1,47) mais l***a concentration finale a un peu augmenté*** (c_{10} passe de 2,78 à 3), c'est à dire les 50% derniers ont toujours à peu prés pareil mais les 10% derniers ont un peu plus ce qui veut dire que les 80% premiers des 50% derniers ont vu leur masse diminuer au profit de celle des 10% derniers.

Remarque :

La comparaison des coefficients de Gini de ces deux séries ne permet évidemment pas d'arriver à une telle conclusion.

9.3.5 Exemple 6

Il s'agit cette fois de la série des revenus des contribuables belges en 1990 (données tirées de la revue Mathématique

et Pédagogie n°104, publiée par la société Belge des Professeurs de mathématique d'expression française), série présentée aussi sous forme non décilée.

x_i est le revenu annuel net en milliers de francs (belges).

n_i est l'effectif correspondant exprimé en milliers de personnes.

Bien entendu x_i est en fait le revenu moyen des n_i individus correspondants.

$n_i x_i$ est exprimé en millions de francs.

x_i	n_i	$n_i x_i$	$\Sigma n_i \searrow$	$\Sigma n_i x_i \searrow$
42,5	200	8500	4107	2826871
153	177	27081	3907	2818371
375	1350	506250	3730	2791290
700	1650	1155000	2380	2285040
1548	730	1130040	730	1130040
	$n = 4107$	$m = 2826871$		

Déterminons le vecteur concentration $\vec{C}(c_{50}, c_{10})$.

$50\%n = 2053,5$ d'où $c_{50} = 2 \times \dfrac{1130040 + (2053,5 - 730) \times 700}{2826871} \simeq 1,46$

$10\%n = 410,7$ d'où $c_{10} = 10 \times \dfrac{410,7 \times 1548}{2826871} \simeq 2,25$

Si on compare avec la série des revenus français 1986 (voir exemple 3) on a une même concentration globale : elle est moyenne, c_{50} étant à peu près égal à 1,5 dans les deux cas ; quant à la concentration finale, faible dans les deux cas, elle est tout de même plus faible pour les belges ($c_{10} \simeq 2,25$) que pour les français ($c_{10} \simeq 2,78$).

Notons là encore, que la comparaison des coefficients de Gini de ces deux séries (0,37 environ pour les deux séries) ne permettrait pas (pour plus de calculs) une telle conclusion.

Enfin $\dfrac{2(c_{50} - 1)}{3} = \dfrac{2 \times 0,46}{3} \simeq 0,31$: on retrouve l'ordre de grandeur du coefficient de Gini (voir 8.3).

9.3.6 Exemple 7

L'objectif de cet exemple est d'illustrer le fait que c_{50} peut être très proche de 1 et pas c_{10}.

Prenons $p = 2$, $x_2 = 2x_1$, $n_1 = 1000$, $n_2 = 100$.

On a $50\%n = 550$ et $10\%n = 110$ d'où

$$c_{50} = 2m_{dr}(0,5) = 2 \times \frac{450x_1 + 100x_2}{1000x_1 + 100x_2} = \frac{1300}{1200} \simeq 1,08$$

$$c_{10} = 10m_{dr}(0,1) = 10 \times \frac{10x_1 + 100x_2}{1000x_1 + 100x_2} = \frac{2100}{1200} \simeq 1,75$$

La concentration globale est donc trés faible puisque 1,08 est trés proche de la valeur 1 caractéristique de la répartition égalitaire.

De même ***la concentration finale*** est trés faible car si 1,75 est plus éloigné de 1 que ne l'est 1,08, 1,75 est tout de même très éloigné de la valeur maximum 10.

Remarque 1 :

On peut constater qu'ici on a rigoureusement $c_{10} = 9c_{50} - 8$: cela est dû au fait que les 90% premiers individus ont la même valeur du caractére puisque $90\%n = 990 < n_1$ (voir 8.2).

Remarque 2 :

Le groupe des 10% derniers a 1,75 fois plus en masse qu'en effectif : on ne peut pas dire que la répartition soit presque égalitaire, c'est à dire que tout groupe a presque autant en masse qu'en effectif, conclusion qu'on aurait tendance à faire si on se contente de calculer le coefficient de Gini (0,08, voir 3.2.2)

9.3.7 Exemple 8

Il s'agit d'étudier ici un cas théorique, celui où ***les valeurs du caractére sont en progression arithmétique et les effectifs correspondants sont constants***.
Prenons comme exemple $n = 10$, $x_i = ia$ pour $i = 1, 2, \ldots, 10$ et avec $a > 0$.
On en déduit tout de suite (pas besoin de tableau de répartition) :

$$c_{50} = 2 \times \frac{(6+7+8+9+10)a}{(1+2+\ldots+10)a} = \frac{16}{11} \simeq 1,45$$
$$c_{10} = 10 \times \frac{10a}{(1+2+\ldots+10)a} = \frac{20}{11} \simeq 1,82 \text{ (qui est le plus fort } mse \text{ de la série)}$$

Donc ***quelque soit la raison*** a, ***la concentration globale*** de cette série est moyenne et sa ***concentration finale*** faible.

Ce résultat n'est pas lié à la valeur 10 de p : on verra en annexe 7 une généralisation au cas $x_i = ia + b$ pour lequel $\lim\limits_{p \to +\infty} c_{50} = 1,5$ et $\lim\limits_{p \to +\infty} c_{10} = 1,9$

Remarque 1 :
Dans le cas étudié ci-dessus ($x_i = ia$) le coefficient de Gini g est égal à 0,3 (appliquer la remarque du 6.5), ce qui ne permet d'obtenir qu'une conclusion beaucoup moins précise : ‹‹concentration›› en dessous de la moyenne.

Remarque 2 :
On peut vérifier (comme au 9.3.1 et 9.3.2) le 8.3, à savoir que $\frac{2}{3}(c_{50} - 1)$ est une bonne valeur approchée de g puisque $\frac{2}{3}(1,45 - 1) = 0,3$. La très bonne qualité de l'approximation étant due ici au fait que la courbe de Lorentz a une allure parabolique (voir début du 6.5).

9.3.8 Exemple 9

Il s'agit ici de reprendre l'exemple du 9.1 dans l'unique but de montrer comment l'exploitation des effectifs et masses en cumulés décroissants et en ***pourcentage*** permet d'obtenir, outre le vecteur concentration, la représentation graphique de la fonction m_{dr}.

x_i	n_i	$n_i x_i$	$\Sigma \frac{n_i}{n} \searrow$	$\Sigma \frac{n_i x_i}{m} \searrow$
5	16	80	1	1
15	30	450	0,8	0,96
30	18	540	0,425	0,75
55	10	550	0,2	0,497
85	6	510	0,075	0,24
	$n = 80$	$m = 2130$		

Il est clair que :
$c_{50} = mse(G_{dr}(0,5)) = \frac{m_{dr}(0,5)}{0,5} = 2 \times (0,75 + (0,5 - 0,425)\frac{0,96 - 0,75}{0,8 - 0,425}) \simeq 1,58$
puisque 0,5 est compris entre 0,8 et 0,425.
La concentration globale est donc est un petit peu au dessus de la moyenne.
$c_{10} = mse(G_{dr}(0,1)) = \frac{m_{dr}(0,1)}{0,1} = 10 \times (0,24 + (0,1 - 0,075)\frac{0,497 - 0,24}{0,2 - 0,075}) \simeq 2,9$
puisque 0,1 est compris entre 0,2 et 0,075.
La concentration finale est donc nettement en dessous de la moyenne.

Cette façon de procéder n'a d'intérêt que si l'on souhaite faire la représentation graphique C_{dr}de la fonction m_{dr}.
En effet $m_{dr}(\alpha)$ = masse (en pourcentage) possédée par les α (en pourcentage) derniers individus et donc les points

(0,075 , 0,24)
(0,2 , 0, 497)
(0,425 , 0,75)
(0,8 , 0,96)

sont sur C_{dr}.

Et comme on a toujours $m_{dr}(0) = 0$ et $m_{dr}(1) = 1$ la courbe C_{dr} est la ligne brisée reliant $O(0,0)$ et $A(1,1)$ et passant par les 4 points ci-dessus (voir 4.1, 4.2).

La voici (c'est la symétrique de la courbe de Lorentz par rapport au point $(\frac{1}{2},\frac{1}{2})$, laquelle a été faite au 6.1 pour ce même exemple) :

Mais à vrai dire cette courbe n'apporte rien de précis par rapport au vecteur concentration $\vec{C}(c_{50},c_{10})$: cependant elle passe évidemment par les points $N_{0,5}(0,5,m_{dr}(0,5) = \frac{c_{50}}{2})$ et $N_{0,1}(0,1,m_{dr}(0,1) = \frac{c_{10}}{10})$ et rappelons le, elle permet de donner une interprétation géométrique de c_{50} et c_{10} car ces deux *mse* ne sont autres que les pentes des droites $(ON_{0,5})$ et $(ON_{0,1})$; voir la remarque du 9.2.1.

Remarque :

L'analyse de cet exemple à l'aide du coefficient de Gini a été faite au 6.1 : on ne peut que conclure à ‹‹concentration›› un peu en dessous de la moyenne puisque $g = 0,427$ est un peu inférieur au milieu de $[0;1[$, cela sans avoir aucune information précise sur les groupes ayant beaucoup plus en masse qu'en effectif.

9.4 Méthode *mse* et méthode Gini-Lorentz

Conclusion

La méthode de Gini Lorentz (détaillée au 3) consiste à faire la courbe de Lorentz et/ou à calculer le coefficient de Gini représentant 2 fois l'aire de la région située entre la courbe de Lorentz et le segment $[OA]$ (égal à la courbe de Lorentz lorsque la répartition est égalitaire).

Si l'interêt de la courbe de Lorentz (ou sa symétrique par rapport au point $(\frac{1}{2},\frac{1}{2})$) est incontestable, car notamment elle est caractéristique de la série à 2 facteurs d'échelle prés, ***le coefficient de Gini, lui, me parait beaucoup moins pertinent***.

Ce coefficient est lourd à calculer et il ne permet d'obtenir qu'une conclusion moins précise que celle de la méthode *mse*. Notamment il ne met en évidence aucun groupe ayant beaucoup plus en masse qu'en effectif (lorsqu'ils existent).

Ceci est dû essentiellement au fait que ***le coefficient de Gini est une valeur moyenne***, en fait égal au double de

la valeur moyenne des différences masse moins effectif : $g = 2\int_0^1 (m_{dr}(\alpha) - \alpha)d\alpha$ (d'aprés 4.5).

Cela a deux conséquences :

1) g est approximativement égal à $\frac{2}{3}(c_{50} - 1)$ (d'aprés 8.3), c'est à dire égal à la première composante c_{50} du vecteur concentration (à une transformation affine près), et donc cela veut dire que calculer le coeffcient de Gini, c'est pratiquement calculer la masse possédée par les 50% derniers individus, cela par un chemin détourné.

Le coefficient de Gini traduit donc surtout ce qui se passe au niveau des 50% derniers individus, cela d'une façon peu parlante.

Il est tout de même ***plus explicite*** de dire que les 50% derniers individus ont un *mse* de 1,58 (1,58 fois plus en masse qu'en effectif) que de dire $g \simeq 0,427$ (voir exemple du 9.3.1) et c'est encore mieux si on compléte le *mse* des 50% derniers individus par le *mse* des 10% derniers, puisque pour un même c_{50} on peut avoir des c_{10} trés différents.

2) En outre le fait que g soit une valeur moyenne explique que ***des séries ayant des courbes de Lorentz trés différentes ont le même*** g, alors que les séries ayant le même vecteur concentration $\vec{C}(c_{50}, c_{10})$ auront des courbes de Lorentz qui passent par 4 mêmes points et donc sans être identiques (il pourra exister des différences notables), ces courbes ne seront pas fondamentalement différentes : par conséquent il en sera de même pour la répartition de la masse.

Enfin la connaissance de ces deux composantes c_{50} et c_{10} permet avec une certaine précision (en tout cas plus qu'avec g) une comparaison entre deux séries ou l'analyse de l'évolution temporelle d'une série (voir 9.3.4).

La méthode *mse* ***peut donc être considérée comme une amélioration du coefficient de Gini*** *:* ***moins de calculs pour parvenir à des conclusions plus précises***.

ANNEXE 1

Propriétés des effectifs et masses en cumulés croissants (pour $p \geq 2$)

1) $\| \frac{\beta_1}{\alpha_1} < \frac{\beta_2}{\alpha_2} <\ldots< \frac{\beta_{p-1}}{\alpha_{p-1}} < 1$
2) $\| \beta_k < \alpha_k$ pour $k \in \{1;2;\ldots;p-1\}$
3) $\| 1 < \frac{1-\beta_1}{1-\alpha_1} < \frac{1-\beta_2}{1-\alpha_2} <\ldots< \frac{1-\beta_{p-1}}{1-\alpha_{p-1}}$
4) $\|$ Si $\bar{x} < x_2$ alors la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est strictement décroissante.
   $\|$ Si $\bar{x} = x_2$ alors la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est strictement décroissante
   $\|$ à partir du rang 2, les deux premiers termes étant égaux.
   $\|$ Si $\bar{x} > x_2$ alors la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est strictement croissante puis
   $\|$ strictement décroissante, avec éventuellement 2 termes égaux à la valeur
   $\|$ maximum de la suite.
   $\|$ La suite $(\alpha_k - \beta_k)_{k\in\{0;1;\ldots;p\}}$ est toujours croissante puis décroissante.
5) $\|$ La valeur maximum de la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est obtenue pour au plus deux
   $\|$ deux valeurs différentes (mais alors consécutives) de k.

Preuve :

1) Pour $k \in \{1;2;\ldots;p\}$ on a $\frac{\beta_k}{\alpha_k} = \frac{n}{m}u_k$ avec $u_k = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i}$.

D'où pour $k \in \{1;2;\ldots;p-1\}$:
$u_k - u_{k+1} = \frac{c_k}{d_k}$ avec $d_k > 0$ et

$$c_k = (\sum_{i=1}^{k} n_i x_i)(\sum_{i=1}^{k} n_i + n_{k+1}) - (\sum_{i=1}^{k} n_i x_i + n_{k+1}x_{k+1})(\sum_{i=1}^{k} n_i).$$
$$= n_{k+1}(\sum_{i=1}^{k} n_i x_i) - n_{k+1}x_{k+1}(\sum_{i=1}^{k} n_i)$$
$$= n_{k+1}(\sum_{i=1}^{k} n_i(x_i - x_{k+1}))$$

Comme la suite $(x_i)_{i\in\{1;2;\ldots;p\}}$ est strictement croissante il vient $x_i - x_{k+1} < 0$ pour $i \in \{1;2;\ldots;k\}$ et donc $c_k < 0$.
Ainsi $\frac{\beta_k}{\alpha_k} - \frac{\beta_{k+1}}{\alpha_{k+1}} < 0$ pour $k \in \{1;2;\ldots;p-1\}$, ce qui prouve le résultat annoncé puisque $\frac{\beta_p}{\alpha_p} = 1$
.

2) C'est une conséquence immédiate du 1 précédent.

3) On procéde comme pour le 1.
Avec la convention $\sum_{i=1}^{0} = 0$, pour $i \in \{0;1;\ldots;p-1\}$ on a :

$$\frac{1-\beta_k}{1-\alpha_k} = \frac{1 - \frac{\sum_{i=1}^{k} n_i x_i}{m}}{1 - \frac{\sum_{i=1}^{k} n_i}{n}} = \frac{n}{m} \times \frac{\sum_{i=k+1}^{p} n_i x_i}{\sum_{i=k+1}^{p} n_i} = \frac{n}{m} \times v_k \text{ avec } v_k = \frac{\sum_{i=k+1}^{p} n_i x_i}{\sum_{i=k+1}^{p} n_i}.$$

D'où pour $k \in \{0;1;2;\ldots;p-2\}$:
$v_k - v_{k+1} = \frac{c_k}{d_k}$ avec $d_k > 0$ et

$$c_k = (\sum_{i=k+2}^{p} n_i)(\sum_{i=k+1}^{p} n_i x_i) - (\sum_{i=k+1}^{p} n_i)(\sum_{i=k+2}^{p} n_i x_i).$$
$$= n_{k+1}x_{k+1}(\sum_{i=k+2}^{p} n_i) - n_{k+1}(\sum_{i=k+2}^{p} n_i x_i)$$
$$= n_{k+1}(\sum_{i=k+2}^{p} n_i(x_{k+1} - x_i))$$

On en déduit $c_k < 0,$ toujours parce que la suite$(x_i)_{i\in\{1;2;\ldots;p\}}$est strictement croissante.
Ainsi $\frac{1-\beta_k}{1-\alpha_k} < \frac{1-\beta_{k+1}}{1-\alpha_{k+1}}$ pour $k \in \{0;1;2;\ldots;p-2\},$ ce qui prouve le résultat annoncé puisque $\frac{1-\beta_0}{1-\alpha_0} = 1.$
Notons que l'inégalité $1 < \frac{1-\beta_1}{1-\alpha_1}$ résulte facilement de $0 < \beta_1 < \alpha_1 < 1$ obtenu en 1.

4) Pour $k \in \{1;2;\ldots;p\}$ on a $\alpha_k - \beta_k = \frac{(\sum_{i=1}^{k} n_i)m - (\sum_{i=1}^{k} n_i x_i)n}{nm} = \frac{\sum_{i=1}^{k} n_i(m - nx_i)}{nm}.$
Etudions la suite $u_k = \sum_{i=1}^{k} n_i(m - nx_i).$
Pour $k \in \{1;2;\ldots;p-1\}$, $u_{k+1} - u_k = n_{k+1}(m - nx_{k+1}) = \frac{n_{k+1}}{n}(\bar{x} - x_{k+1}).$
Rappelons que d'après 2.2.3 on a $\bar{x} \in]x_1;x_p[$ (on est dans le cas $p \geq 2$).
Tous les effectifs étant strictement positifs, le sens de variation de la suite ne dépend que du signe des $\bar{x} - x_{k+1}$.
Compte tenu toujours du fait que la suite $(x_i)_{i\in\{1;2;\ldots;p\}}$ est strictement croissante, 3 possibilités se présentent :
si $\bar{x} < x_2$ alors $\forall\ k \in \{1;2;\ldots;p-1\}\ u_{k+1} < u_k$ et la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est strictement décroissante (à partir du rang 1).
si $\bar{x} = x_2$ (donc $p \geq 3$ puisque $\bar{x} \in]x_1;x_2[$ si $p = 2$) alors $u_1 = u_2$ et $\forall\ k \in \{2;\ldots;p-1\}\ u_{k+1} < u_k$ et la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est strictement décroissante à partir du rang 2.
si $\bar{x} > x_2$ (donc $p \geq 3$, même raison que ci-dessus), comme $\bar{x} < x_p$ il existe $k_0 \in \{2;3;\ldots;p-1\}$ tel que :
$\bar{x} - x_2 > 0,\ \bar{x} - x_3 > 0,\ldots,\ \bar{x} - x_{k_0} > 0$
$\bar{x} - x_{k_0+1} \leq 0$ et si $k_0 \leq p-2$ alors $\bar{x} - x_{k_0+2} < 0, \bar{x} - x_{k_0+3} < 0,\ldots,\ \bar{x} - x_p < 0.$
D'où $u_1 < u_2 <\ldots< u_{k_0}$ et $u_{k_0} \geq u_{k_0+1} > u_{k_0+2} >\ldots> u_p.$
La suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ est donc d'abord strictement croissante puis strictement décroissante, avec éventuellement deux termes égaux à la valeur maximum de la suite $(u_{k_0} = u_{k_0+1} \Leftrightarrow \bar{x} = x_{k_0+1})$.
Notons que k_0 peut être égal à $p-1,$ auquel cas cette suite est strictement croissante jusqu'au rang $p-1$ et ensuite elle est bien strictement décroissante puisque $\alpha_p - \beta_p = 0$: voir exemple 4 du 6.3.
Enfin compte tenu de $\alpha_0 - \beta_0 = 0$ et de $\alpha_1 - \beta_1 > 0$ on peut évidemment dire que la suite $(\alpha_k - \beta_k)_{k\in\{0;1;\ldots;p\}}$ est toujours croissante puis décroissante.

5) Conséquence évidente du 4. Précisons un tout petit peu.
On note $e = \max_{k\in\{0;1;\ldots;p\}} \alpha_k - \beta_k = \max_{k\in\{1;2;\ldots;p-1\}} \alpha_k - \beta_k,$ car $\alpha_0 - \beta_0 = \alpha_p - \beta_p = 0$ et $\alpha_1 - \beta_1 > 0.$
Si $\bar{x} < x_2$ alors $e = \alpha_k - \beta_k$ uniquement pour $k = 1$
Si $\bar{x} = x_2$ alors $e = \alpha_k - \beta_k$ pour $k = 1$ et pour $k = 2$
Si $\bar{x} > x_2$ alors $e = \alpha_k - \beta_k$ uniquement pour $k = k_0$ si $\bar{x} < x_{k_0+1}$ et uniquement pour $k = k_0$ et $k = k_0 + 1$ si $k_0 + 1 < p$ et $\bar{x} = x_{k_0+1}$.

Remarque :
Donnons deux exemples :
Exemple avec $\bar{x} < x_2$

x_i	n_i	$n_i x_i$	α_i	β_i	$\alpha_i - \beta_i$
5	996	4980	0,996	0,9651	0,0309
15	1	15	0,997	0,9680	0,029
30	1	30	0,998	0,9738	0,0242
50	1	50	0,999	0,9835	0,0155
85	1	85	1	1	0
	$n = 1000$	$m = 5160$			

On a $\bar{x} = 5,160 < x_2$ et la suite $(\alpha_k - \beta_k)_{k\in\{1;2;3;4;5\}}$ est bien strictement décroissante.

Exemple avec $\bar{x} > x_2$

x_i	n_i	$n_i x_i$	α_i	β_i	$\alpha_i - \beta_i$
5	200	1000	12/60	2/60	10/60
15	300	4500	30/60	11/60	19/60
30	200	6000	42/60	23/60	19/60
50	200	10000	54/60	43/60	11/60
85	100	8500	1	1	0
	$n = 1000$	$m = 30000$			

On a $\bar{x} = 30 > x_2$ avec la particularité supplémentaire $\bar{x} = x_3$: la suite $(\alpha_k - \beta_k)_{k\in\{1;2;3;4;5\}}$ est bien strictement croissante puis strictement décroissante avec deux termes égaux ($k_0 = 2$).

ANNEXE 2

Propriété de la courbe de Lorentz C_{pr} (pour $p \geq 2$)

1) ‖ Pour $k \in \{0;1;\ldots;p-1\}$:
   ‖ la pente de la droite(OM_k) est inférieure à la pente de la droite (OM_{k+1})
2) ‖ C_{pr} est strictement en dessous de $[OA]$ sauf les points $M_0 = O$ et $M_p = A$
3) ‖ Les pentes des segments $[M_kM_{k+1}]$, respectivement égales à $\frac{x_{k+1}}{\bar{x}}$, forment une suite
   ‖ strictement croissante pour $k \in \{0;1;\ldots;p-1\}$
4) ‖ C_{pr} admet exactement $p-1$ points anguleux.
5) ‖ C_{pr} est la représentation graphique d'une fonction notée m_{pr} qui est affine par intervalle
   ‖ et convexe. De façon plus précise on a :
   ‖ pour $k \in \{0;1;\ldots;p-1\}$, $m_{pr}(\alpha) = a_k\alpha + b_k$ sur $[\alpha_k;\alpha_{k+1}]$
   ‖ avec $a_k = \frac{x_{k+1}}{\bar{x}}$ et $b_k = \beta_k - \frac{x_{k+1}}{\bar{x}}\alpha_k = \beta_{k+1} - \frac{x_{k+1}}{\bar{x}}\alpha_{k+1}$
   ‖ Les coefficients a_k et b_k vérifient $b_0 = 0$ et $b_{k-1} < 0$ pour $1 \leq k \leq p-1$
   ‖ ainsi que $a_{p-1} + b_{p-1} = 1$ et $0 < a_k + b_k < 1$ pour $0 \leq k \leq p-2$
6) ‖ Si $1 \leq k \leq p-1$ la droite (M_kM_{k+1}) coupe l'axe des ordonnées en dessous de O
   ‖ Si $0 \leq k \leq p-2$ la droite (M_kM_{k+1}) coupe la droite (AB) entre A et B.
7) ‖ Les points M_k s'éloignent d'abord de $[OA]$ puis s'en rapprochent.

Preuve :

1) La pente de la droite (OM_k) est $\frac{\beta_k}{\alpha_k}$, d'où le résultat d'après le 1 de l'annexe 1.

2) C'est la traduction géométrique de l'inégalité $\beta_k < \alpha_k$ obtenue au 2 de l'annexe 1.

3) La pente du segment $[M_kM_{k+1}]$ est $\frac{\beta_{k+1}-\beta_k}{\alpha_{k+1}-\alpha_k} = \frac{\frac{n_{k+1}x_{k+1}}{m}}{\frac{n_{k+1}}{n}} = \frac{x_{k+1}}{\bar{x}}$;
Compte tenu de la stricte croissance de la suite $(x_i)_{i\in\{1;2;\ldots;p\}}$ on obtient le résultat annoncé.

4) C'est une conséquence du point 3 précédent.

5) C_{pr} étant la ligne brisée reliant les points M_k, la fonction m_{pr} est affine sur chaque intervalle $[\alpha_k;\alpha_{k+1}]$, c'est à-dire sur cet intervalle $m_{pr}(\alpha) = a_k\alpha + b_k$.
a_k est la pente du segment $[M_kM_{k+1}]$ et donc $a_k = \frac{x_{k+1}}{\bar{x}}$; et puisque $m_{pr}(\alpha_k) = \beta_k$ on a
$b_k = \beta_k - \frac{x_{k+1}}{\bar{x}}\alpha_k = \beta_{k+1} - \frac{x_{k+1}}{\bar{x}}\alpha_{k+1}$.
On en déduit $b_0 = \beta_0 - \frac{x_1}{\bar{x}}\alpha_0 = 0$ (car β_0 et α_0 sont nuls, ce qui correspond au fait que $M_0 = O$).
Pour $1 \leq k \leq p-1$:
$$b_k = \frac{\sum_{i=1}^{k} n_i x_i}{m} - \frac{n}{m}x_{k+1}\frac{\sum_{i=1}^{k} n_i}{n} = \frac{\sum_{i=1}^{k} n_i(x_i - x_{k+1})}{m}.$$

Ce qui prouve que b_k est négatif, la suite $(x_i)_{i\in\{1;2;\ldots;p\}}$ étant strictement croissante.
Pour $0 \leq k \leq p-1$:
$a_k + b_k = a_k + \beta_k - a_k\alpha_k = a_k(1-\alpha_k) + \beta_k$ et comme $a_k > 0$ et $1 - \alpha_k > 0$ (car k est ici différent de p) on obtient $a_k + b_k > 0$.
En fait pour $k = p-1$ on peut préciser la valeur de $a_{p-1} + b_{p-1}$: soit on calcule $a_{p-1}(1-\alpha_{p-1}) + \beta_{p-1}$ et on trouve 1, soit on remarque que 1 est dans $[\alpha_{p-1};\alpha_p] = [\alpha_{p-1};1]$ et donc on peut écrire $m_{pr}(1) = a_{p-1} \times 1 + b_{p-1}$, ce qui donne $a_{p-1} + b_{p-1} = 1$ (puisque $m_{pr}(1) = 1$).

Pour $0 \leq k \leq p-2$, et en tenant compte de la valeur trouvée ci-dessus pour b_k :

$$\begin{aligned} 1-a_k-b_k &= 1-\frac{n}{m}x_{k+1}-\frac{\sum_{i=1}^{k} n_i(x_i-x_{k+1})}{m} \\ &= \frac{m-\sum_{i=1}^{k} n_i x_i+(\sum_{i=1}^{k} n_i)x_{k+1}-nx_{k+1}}{m} \\ &= \frac{\sum_{i=k+1}^{p} n_i x_i-(\sum_{i=k+1}^{p} n_i)x_{k+1}}{m} \\ &= \frac{\sum_{i=k+1}^{p} n_i(x_i-x_{k+1})}{m} \\ &= \frac{\sum_{i=k+2}^{p} n_i(x_i-x_{k+1})}{m} \end{aligned}$$

et donc $1-a_k-b_k > 0,$ toujours parce que la suite $(x_i)_{i\in\{1;2;\ldots;p\}}$ est strictement croissante.

Montrons la convexité de la fonction m_{pr} : en tout point de $[\alpha_k;\alpha_{k+1}[$ m_{pr} admet une dérivée droite égale à $\frac{x_{p+1}}{x},$ donc sur $[0;1[$ la dérivée droite de m_{pr} est croissante et m_{pr} est convexe sur $[0;1[,$ et par continuité elle est convexe sur $[0;1]$.

6) Le premier point résulte de $b_k < 0$ et le deuxième point résulte de $0 < a_k + b_k < 1$ car le point d'abscisse 1 de la droite (M_kM_{k+1}) a pour ordonnée $a_k + b_k$.

7) Au cours de la preuve du 3.2.1 (illustrée d'une figure) on a montré que la distance du point $M_k(\alpha_k, \beta_k)$ au segment $[OA]$ est $\frac{\sqrt{2}}{2}(\alpha_k - \beta_k)$; la suite $(\alpha_k - \beta_k)_{k\in\{1;2;\ldots;p\}}$ étant d'abord croissante puis décroissante (voir annexe 1) on obtient alors le résultat annoncé.

ANNEXE 3

Majoration du coefficient de Gini

1) ‖ $g \leq \dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$

2) ‖ Pour $p \geq 2$

‖ $g = \dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}} \Leftrightarrow p = 2$ et $\bar{x} = \sqrt{x_1 x_2}$ ($\bar{x}$ est la moyenne géométrique de x_1 et x_2)

‖ $\Leftrightarrow p = 2$ et $\dfrac{n_1}{n} = \dfrac{\sqrt{x_2}}{\sqrt{x_2} + \sqrt{x_1}}$

‖ On verra à la remarque 2 ci-après que g peut être très proche de $\dfrac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}}$

‖ sans que pour autant p soit égal à 2.

Preuve :

Pour l'instant, je garde la preuve de ces réultats pour moi.

.

ANNEXE 4

Effet d'une translation sur le coefficient de Gini et sur les mse

‖ Si on augmente d'un même nombre k toutes les valeurs du caractère d'une série alors :

1) ‖ Le coefficient de Gini g de la série devient $g' = \dfrac{g}{1 + k\frac{n}{m}}$

2) ‖ Le *mse* de n'importe quel sous-groupe G d'individus de la série devient

‖ $mse'(G) = mse(G)\dfrac{1 + \frac{k}{\bar{x}_G}}{1 + k\frac{n}{m}}$

Preuve :

1) Si $p = 1$ alors $g = g' = 0$ et la relation annoncée est évidemment vraie.

On suppose maintenant $p \geq 2$.

Cf le 3.1.4 on a $g = 1 - \sum_{i=0}^{p-1} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1}) = 1 - \sum_{i=0}^{p-1} f_{i+1}(\beta_i + \beta_{i+1})$, avec $\beta_0 = 0$ et pour $i \geq 1$

$\beta_i = \dfrac{\sum_{j=1}^{i} n_j x_j}{m} = \frac{n}{m}\sum_{j=1}^{i} f_j x_j$.

D'où $\frac{m}{n}g = \frac{m}{n} - \sum_{i=0}^{p-1} f_{i+1}(\sum_{j=1}^{i} f_j x_j + \sum_{j=1}^{i+1} f_j x_j)$, en convenant que $\sum_{j=1}^{i} f_j x_j = 0$ pour $i = 0$.

Si on augmente de k toutes les valeurs du caractère, alors la moyenne $\frac{m}{n}$ augmente aussi de k et la relation précédente permet d'écrire que le nouveau coefficient de Gini g' est donné par la relation :

$(\frac{m}{n} + k)g' = \frac{m}{n} + k - \sum_{i=0}^{p-1} f_{i+1}(\sum_{j=1}^{i} f_j x_j + \sum_{j=1}^{i+1} f_j x_j + k(\sum_{j=1}^{i} f_j + \sum_{j=1}^{i+1} f_j)) = \frac{m}{n}g + k - ks$ avec

$s = \sum_{i=0}^{p-1} f_{i+1}(\sum_{j=1}^{i} f_j + \sum_{j=1}^{i+1} f_j) = \sum_{i=0}^{p-1} f_{i+1}^2 + 2\sum_{i=0}^{p-1} f_{i+1}(\sum_{j=1}^{i} f_j) = \sum_{i=0}^{p-1} f_{i+1}^2 + 2\sum_{i=1}^{p-1} f_{i+1}(\sum_{j=1}^{i} f_j) = \sum_{i=1}^{p} f_i^2 + 2\sum_{i=2}^{p} f_i(\sum_{j=1}^{i-1} f_j)$

$s = \sum_{i=1}^{p} f_i^2 + 2(f_2 f_1 + f_3(f_1 + f_2) + f_4(f_1 + f_2 + f_3) + \ldots + f_p(f_1 + f_2 + \ldots + f_{p-1})$

$s = \sum_{i=1}^{p} f_i^2 + 2(f_1(f_2 + f_3 + \ldots + f_p) + f_2(f_3 + f_4 + \ldots + f_p) + \ldots f_{p-1}f_p)$

$s = \sum_{i=1}^{p} f_i^2 + 2\sum_{1 \leq i < j \leq p}^{p} f_i f_j = (f_1 + f_2 + \ldots + f_p)^2 = 1^2 = 1$

et ainsi $(\frac{m}{n} + k)g' = \frac{m}{n}g$ ce qui donne le résultat annoncé.

2) Cf 2.2.2 on sait que $mse(G) = \frac{\bar{x}_G}{\bar{x}}$ avec $\bar{x} = \frac{m}{n}$ moyenne de la série et $\bar{x}_G$ moyenne du groupe G ; donc

$mse'(G) = \dfrac{\bar{x}_G + k}{\bar{x} + k} = \dfrac{\bar{x}_G(1 + \frac{k}{\bar{x}_G})}{\bar{x}(1 + \frac{k}{\bar{x}})}$, ce qui donne le résultat annoncé.

Remarque 1 :

$\lim_{k\to+\infty} g' = 0$; résultat attendu puisque si k devient très très grand tous les individus ont "relativement" la même valeur du caractère donc la répartition est quasiment égalitaire.

Remarque 2 :

$\lim_{k\to+\infty} mse'(G) = 1$; résultat là aussi attendu : voir 2.2.3.

On a aussi : $mse'(G) > mse(G) \Leftrightarrow \bar{x}_G < \bar{x}$ et $mse'(G) = mse(G) \Leftrightarrow \bar{x}_G = \bar{x}$.

Remarque 3 :

Dans l'exemple 7 du 6.5 on a vu que $g = \frac{1}{3} \times \frac{1-\frac{1}{p^2}}{1+\frac{a+2b}{pa}}$ et donc $g' = \frac{1}{3} \times \frac{1-\frac{1}{p^2}}{1+\frac{a+2(b+k)}{pa}}$; vérifions la relation trouvée au 1) ci-dessus :

$\frac{g'}{g} = \frac{pa+a+2b}{pa+a+2(b+k)} = \frac{\frac{(p+1)a}{2}+b}{\frac{(p+1)a}{2}+b+k}$;

mais $m = n_1x_1 + \ldots + n_px_p = \frac{n}{p}(x_1 + \ldots + x_p) = \frac{n}{p}(pb + \frac{p(p+1)a}{2})$, soit $m = n(b + \frac{(p+1)a}{2})$ et on obtient $\frac{g'}{g} = \frac{\frac{m}{n}}{\frac{m}{n}+k} = \frac{1}{1+\frac{kn}{m}}$.

ANNEXE 5

Un résultat étonnant sur le coefficient de Gini

‖ Soit X la variable aléatoire égale à la valeur absolue de la différence des valeurs du caractère
‖ de deux individus choisis au hasard parmi les n de la population. Précisons ce choix :
‖ on choisit au hasard un individu parmi les n, puis on choisit encore au hasard un individu
‖ parmi les n, c'est-à-dire on répéte deux fois (avec indépendance) la même épreuve :
‖ choisir un individu parmi les n.
‖ Alors l'espérance mathématique de X, ou valeur moyenne, est le double du coefficient de Gini
‖ multiplié par la masse moyenne, soit $E(X) = 2g\frac{m}{n}$
‖ *On a ici une 2ième intreprétation de g en terme de valeur moyenne de différences,*
‖ *la 1ière interprétation étant celle du 1) et 2) de la propriété 4.5*

Preuve :

1) Preuve dans le cas particulier de la remarque 1 du 3.2.3.

On a $p = 2,\ x_1 = 0,\ x_2 > 0,\ n_1 = n-1,\ n_2 = 1$ et $g = \frac{n-1}{n}$ et donc $2g\frac{m}{n} = \frac{2(n-1)m}{n^2}$.

Ici X ne prend que 2 valeurs : 0 et m (puisque $m = n_1x_1 + n_2x_2 = x_2$).

$P(X = 0)$ est la probabilité que les deux individus choisis aient tous les deux 0 ou tous les deux m et donc $P(X = 0) = \frac{n-1}{n} \times \frac{n-1}{n} + \frac{1}{n} \times \frac{1}{n}$ (en fait ce résultat est inutile pour le calcul de l'espérance).

$P(X = 1)$ est la probabilité que l'un des individus a 0 et l'autre m et $P(X = 1) = 2\frac{n-1}{n} \times \frac{1}{n}$.

D'où $E(X) = 2\frac{n-1}{n} \times \frac{1}{n} \times m = 2g\frac{m}{n}$.

2) Preuve dans le cas particulier correspondant à l'exemple 7 du 6.5.

On a $x_i = ia + b$ pour $i = 1, 2, \ldots, p$ et les n_i sont tous égaux à $\frac{n}{p}$.

On a $g = \frac{a(p^2-1)}{6p(\frac{p+1}{2}a + b)}$ (voir 6.5) et $\frac{m}{n} = \frac{n_1x_1 + \ldots + n_px_p}{n} = \frac{p+1}{2}a + b$ (voir remarque 3 de l'annexe 4) et donc $2g\frac{m}{n} = \frac{a(p^2-1)}{3p}$.

Cette fois X prend les valeurs $0, a, 2a, 3a, \ldots, (p-1)a$.

Notons tout de suite que dans le cas particulier où $p = 1$ alors $g = 0$ et X ne prenant que la valeur 0, $E(X) = 0$ et on a bien $E(X) = 2g\frac{m}{n}$.

Prouvons cela pour p quelconque (≥ 2).

$P(X = 0) = p \times \frac{1}{p^2}$ car p tirages donnent 0 : $(x_1, x_1), (x_2, x_2), \ldots, (x_p, x_p)$ et chacun de ces tirages a comme probabilité le produit des fréquences correpondantes d'apparition soit $\frac{\frac{n}{p}}{n} \times \frac{\frac{n}{p}}{n} = \frac{1}{p^2}$.

Pour $i = 1, 2, \ldots, p-1$ on a $P(X = ia) = 2\frac{p-i}{p^2}$ car il y a $2(p-i)$ tirages qui donnent ia, les $p-i$ tirages suivants $(x_1, x_{1+i}), (x_2, x_{2+i}), \ldots, (x_{p-i}, x_p)$ et les symétriques, lesquels ont encore tous $\frac{1}{p^2}$ comme probabilité d'apparition.

On a alors $E(X) = \sum_{i=1}^{p-1} \frac{2(p-i)}{p^2} ia = \frac{2a}{p^2}(p\sum_{i=1}^{p-1} i - \sum_{i=1}^{p-1} i^2) = \frac{2a}{p^2}(p\frac{(p-1)p}{2} - \frac{(p-1)p(2p-1)}{6})$, soit

$E(X) = \frac{a(p-1)}{p}(p - \frac{2p-1}{3}) = \frac{a(p^2-1)}{3p} = 2g\frac{m}{n}$.

3) Preuve dans le cas général.

Cf le 3.1.4 on a $g = 1 - \sum_{i=0}^{p-1} \frac{n_{i+1}}{n}(\beta_i + \beta_{i+1}) = 1 - \sum_{i=0}^{p-1} f_{i+1}(\beta_i + \beta_{i+1})$, avec $\beta_0 = 0$ et pour $i \geq 1$

$\beta_i = \frac{\sum_{j=1}^{i} n_jx_j}{m} = \frac{n}{m}\sum_{j=1}^{i} f_jx_j$.

D'où $2\frac{m}{n}g = 2\frac{m}{n} - 2\sum_{i=0}^{p-1} f_{i+1}(\sum_{j=1}^{i} f_j x_j + \sum_{j=1}^{i+1} f_j x_j)$, en convenant que $\sum_{j=1}^{i} f_j x_j = 0$ pour $i = 0$; soit, en remplacant i par $i-1$, $2\frac{m}{n}g = 2\frac{m}{n} - 2\sum_{i=2}^{p} f_i(\sum_{j=1}^{i-1} f_j x_j) - 2\sum_{i=1}^{p} f_i(\sum_{j=1}^{i} f_j x_j)$.

Les valeurs prises par X sont $|x_i - x_j|$ (voir remarque 1) pour $(i,j) \in \{1;2;\ldots;p\}^2$ avec $P(X = |x_i - x_j|) = f_i f_j$; on peut évidemment vérifier que la somme de ces probabilités est bien 1, puisque $\sum_{\substack{1\le i\le p\\ 1\le j\le p}} f_i f_j = (\sum f_i)(\sum f_j) = 1 \times 1$.

$E(X) = \sum_{i=1}^{p}\sum_{j=1}^{p} f_i f_j |x_i - x_j| = 2\sum_{i=2}^{p}(\sum_{j=1}^{i-1} f_i f_j (x_i - x_j))$, car pour $i = j$ on a une contribution nulle et (i,j) et (j,i) donnent la même chose, donc on peut considérer que les couples (i,j) avec $1 \le j \le i-1$ et $i \ge 2$.

Ainsi $E(X) = 2\sum_{i=2}^{p} f_i(x_i(\sum_{j=1}^{i-1} f_j) - \sum_{j=1}^{i-1} f_j x_j)$.

D'où, montrer que $E(X) = 2\frac{m}{n}g$ revient à montrer que (voir plus haut l'expression de $2\frac{m}{n}g$)

$\sum_{i=2}^{p} f_i(x_i(\sum_{j=1}^{i-1} f_j) - \sum_{j=1}^{i-1} f_j x_j) + \sum_{i=2}^{p} f_i(\sum_{j=1}^{i-1} f_j x_j) + \sum_{i=1}^{p} f_i(\sum_{j=1}^{i} f_j x_j) = \frac{m}{n}$ soit

$\sum_{i=2}^{p} f_i x_i(\sum_{j=1}^{i-1} f_j) + \sum_{i=1}^{p} f_i(\sum_{j=1}^{i} f_j x_j) = \sum_{i=1}^{p} f_i x_i$.

Le coefficient de x_i dans la 1ère partie du membre de gauche est évidemment $f_i \sum_{j=1}^{i-1} f_j$

En écrivant ligne par ligne les termes de la 2ième partie du membre de gauche on obtient :

$f_1(f_1 x_1)$

$f_2(f_1 x_1 + f_2 x_2)$

$f_3(f_1 x_1 + f_2 x_2 + f_3 x_3)$

....

$f_p(f_1 x_1 + f_2 x_2 + f_3 x_3 + \ldots + f_p x_p)$.

On voit alors clairement que le coeffcient de x_i dans cette 2ième partie est $f_i(\sum_{j=i}^{p} f_j)$ et finalement le coeffcient de x_i dans le membre de gauche est $f_i(\sum_{j=1}^{i-1} f_j + \sum_{j=i}^{p} f_j) = f_i$, qui est le coefficient de x_i dans le membre de droite et ainsi l'égalité est prouvée!

Remarque 1 :

En fait les $|x_i - x_j|$ ne sont pas tous distincts ; par exemple si $i = j$ on obtient p fois la valeur 0, et donc une démonstration propre exigerait de ne considérer que les valeurs distinctes $v_1, v_2, \ldots v_k$ prises par X. Dans ce cas $P(X = v_l)$ serait égal à la somme des $f_i f_j$, pour tous les couples (i,j) tels que $|x_i - x_j| = v_l$ et .... on retrouverait la même expression pour $E(X)$.

Remarque 2 :

Ce résultat permet de retrouver rapidement celui de l'annexe 4 sur l'effet d'une translation sur le coefficient de Gini.

Soit g le coefficient de Gini d'une série (x_i, n_i) et g' le coefficient de la série obtenue en augmentant tous les x_i d'une même quantité k, les effectifs n_i restant inchangés.

L'espérance de X reste inchangée mais la valeur moyenne de la série devient $\frac{m}{n} + k$ d'où $2g\frac{m}{n} = 2g'(\frac{m}{n} + k)$, ce qui redonne $g' = \frac{g}{1 + k\frac{m}{n}}$.

ANNEXE 6

Valeurs moyennes des mse

1) ‖ Dans tout ce qui suit les fonctions $\alpha \to mse(G_{dr}(\alpha))$ et $\alpha \to mse(G_{pr}(\alpha))$
‖définies sur $]0;1]$ seront prolongées par continuité en 0 en posant $mse(G_{dr}(0)) = \frac{x_p}{\bar{x}}$
‖ et $mse(G_{pr}(0)) = \frac{x_1}{\bar{x}}$
‖ Ainsi :
‖ $\forall k \in \{0;1;\ldots;p-1\}$ sur $[\alpha_k;\alpha_{k+1}]$ on a $mse(G_{pr}(\alpha)) = a_k + \frac{b_k}{\alpha}$
‖ avec $a_k = \frac{x_{k+1}}{\bar{x}}$ et $b_k = \beta_k - a_k\alpha_k$ $(b_0 = 0)$
‖ $\forall k \in \{0;1;\ldots;p-1\}$ sur $[1-\alpha_{k+1};1-\alpha_k]$ on a $mse(G_{dr}(\alpha)) = a_k + \frac{1-a_k-b_k}{\alpha}$
‖ $(1 - a_0 - b_0 = 0)$

2) ‖ Un exemple de représentation graphique de la fonction $\alpha \to mse(G_{dr}(\alpha))$ sur $[0;1]$
‖ Cette courbe est toujours constituée d'un palier horizontal et de $p-1$ arcs d'hyperboles.

3) ‖ $K_p = \int_0^1 mse(G_{dr}(\alpha))d\alpha \in \left[1; \frac{x_p}{\bar{x}}\right] \subset [1;+\infty[$

4) ‖ $K_p = 1 \Leftrightarrow K_p = \frac{x_p}{\bar{x}} \Leftrightarrow$ la répartition est égalitaire (c'est-à-dire $p = 1$)

5) ‖ Si $p \geq 2$ alors $K_p = 1 + \sum_{i=0}^{p-2}(1 - \beta_i - \frac{x_{i+1}}{\bar{x}}(1-\alpha_i)) \ln\frac{1-\alpha_i}{1-\alpha_{i+1}}$
‖ en particulier $K_2 = 1 - \frac{\bar{x}-x_1}{\bar{x}}\ln(1-\frac{n_1}{n})$

6) ‖ Si $p = 2$ (donc $n \geq 2$), $x_1 = 0$, $n_2 = 1$ (le dernier individu a tout) alors $K_2 = 1 + \ln n$

7) ‖ Un exemple de calcul de K_p

8) ‖ $J_p = \int_0^1 mse(G_{pr}(\alpha))d\alpha \in \left[\frac{x_1}{\bar{x}};1\right] \subset [0;1]$

9) ‖ $J_p = 1 \Leftrightarrow J_p = \frac{x_1}{\bar{x}} \Leftrightarrow$ la répartition est égalitaire (c'est-à-dire $p = 1$)

10) ‖ Si $p \geq 2$ alors $J_p = 1 + \sum_{i=1}^{p-1}(\beta_i - \frac{x_{i+1}}{\bar{x}}\alpha_i)\ln\frac{\alpha_{i+1}}{\alpha_i}$
‖ en particulier $J_2 = 1 - \frac{n_1}{n}\frac{x_1-x_2}{\bar{x}}\ln\frac{n_1}{n}$

11) ‖Si $p = 2$ (donc $n \geq 2$), $x_1 = 0$, $n_2 = 1$ (le dernier individu a tout) alors
‖ $J_2 = 1 + (n-1)\ln(1-\frac{1}{n})$, quantité équivalente à $\frac{1}{2n}$ pour n grand.

Preuve :

1) Si $p = 1$, d'après le 1 de 7.2 (qui est en fait une conséquence immédiate du 1 de 2.2.3) on sait que $\forall\alpha \in]0;1]$ $mse(G_{pr}(\alpha)) = mse(G_{dr}(\alpha)) = 1$: ces deux fonctions sont donc prolongeables par continuité en 0 en posant $mse(G_{pr}(0)) = mse(G_{dr}(0)) = 1 = \frac{x_1}{\bar{x}} = \frac{x_p}{\bar{x}}$.

Pour $p \geq 2$ on va reprendre une partie des calculs faits lors de la preuve du 2 et 3 de 7.2.

D'après l'annexe 2, $\forall\ k \in \{0;1;\ldots;p-1\}$, sur $[\alpha_k;\alpha_{k+1}]$ $m_{pr}(\alpha) = a_k\alpha + b_k$
avec $a_k = \frac{x_{k+1}}{\bar{x}}$ et $b_k = \beta_k - \frac{x_{k+1}}{\bar{x}}\alpha_k = \beta_{k+1} - \frac{x_{k+1}}{\bar{x}}\alpha_{k+1}$ et aussi $b_0 = 1 - a_{p-1} - b_{p-1} = 0$.

Donc sur $]0;\alpha_1]$ $mse(G_{pr}(\alpha)) = a_0 + \frac{b_0}{\alpha} = a_0 = \frac{x_1}{\bar{x}}$: on peut prolonger la fonction $\alpha \to mse(G_{pr}(\alpha))$ par continuité en 0 en posant $mse(G_{pr}(0)) = \frac{x_1}{\bar{x}}$.

Ainsi $\forall k \in \{0;1;\ldots;p-1\}$, sur $[\alpha_k;\alpha_{k+1}]$ on a $mse(G_{pr}(\alpha)) = a_k + \frac{b_k}{\alpha}$ (vrai aussi dans le cas $p = 1$).

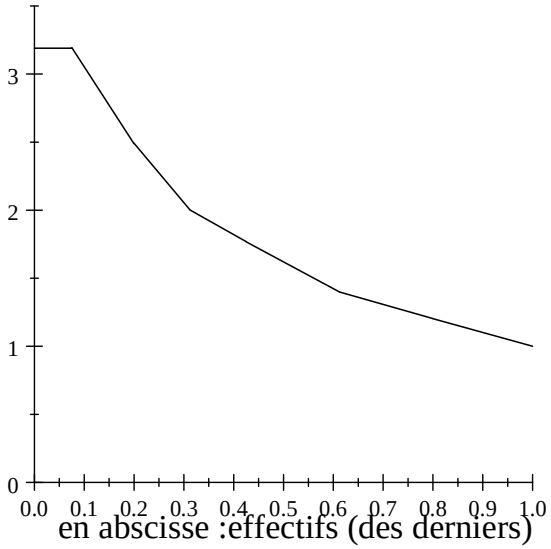
De $m_{dr}(\alpha) = 1 - m_{pr}(1-\alpha)$ on déduit que sur $[1-\alpha_{k+1};1-\alpha_k]$ $m_{dr}(\alpha) = a_k\alpha + 1 - a_k - b_k$ et donc sur $]0;1-\alpha_{p-1}]$ $mse(G_{dr}(\alpha)) = a_{p-1} + \frac{1-a_{p-1}-b_{p-1}}{\alpha} = a_{p-1} = \frac{x_p}{\bar{x}}$:
on peut donc prolonger la fonction $\alpha \to mse(G_{dr}(\alpha))$ par continuité en 0 en posant $mse(G_{dr}(0)) = \frac{x_p}{\bar{x}}$.

Ainsi $\forall k \in \{0;1;\ldots;p-1\}$ sur $[1-\alpha_{k+1};1-\alpha_k]$ on a $mse(G_{dr}(\alpha)) = a_k + \frac{1-a_k-b_k}{\alpha}$ (vrai aussi dans le cas $p = 1$)

2) Pour obtenir la représentation graphique de $\alpha \to mse(G_{dr}(\alpha))$ il suffit d'utiliser les formules du 1. Illustrons cela avec l'exemple du 6.1 ($p = 5$) :

k	α_k	β_k	a_k	b_k	$1 - a_k - b_k$	sur	$mse(G_{dr}(\alpha))$
0	0	0	0,1878	0	0,8122	$[0,8;1]$	$0,1878 + \frac{0,8122}{\alpha}$
1	0,2	0,0375	0,5634	-0,07518	0,51178	$[0,425;0,8]$	$0,5634 + \frac{0,51178}{\alpha}$
2	0,575	0,248	1,1267	-0,3998	0,2731	$[0,2;0,425]$	$1,1267 + \frac{0,2731}{\alpha}$
2	0,8	0,502	2,0657	-1,1505	0,0848	$[0,075;0,2]$	$2,0657 + \frac{0,0848}{\alpha}$
4	0,925	0,760	3,1924	-2,1929	0	$[0;0,075]$	3,1924

ce qui donne comme représentation graphique :



Courbe de $\alpha \to mse(G_{dr}(\alpha))$

La courbe est donc constituée d'un palier horizontal (dont la hauteur, $\frac{x_5}{\bar{x}} = 3,19$, est le plus fort *mse* de la série) et de 4 arcs d'hyperboles ; la longueur du palier horizontal est la fréquence du caractère le plus élevè (x_5), la longueur de l'arc d'hyperbole immédiatement après est la fréquence du caractère x_4, et ainsi de suite.
Le graphique permet évidemment de retrouver le sens de variation de cette fonction (voir le 3 de 7.2).

3) $mse(G_{pr}(\alpha)) \in \left[1; \frac{x_p}{\bar{x}}\right]$ pour tout α dans$[0;1]$ (d'après 7.4 et le fait que $mse(G_{dr}(0)) = \frac{x_p}{\bar{x}}$), ce qui prouve le résultat demandé puisque l'intégration se fait sur un intervalle de longueur 1.

4) Si $p = 1$ alors $\forall\ \alpha \in [0;1]$ on a $mse(G_{dr}(\alpha)) = 1$ et donc $K_p = 1 = \frac{x_p}{\bar{x}}$.
Si $K_p = 1$ c'est que $\int_0^1 (mse(G_{dr}(\alpha)) - 1)d\alpha = 0$, et comme la fonction à intégrer est continue, positive ou nulle, obligatoirement elle est nulle partout et donc $\forall \alpha \in [0;1]$ on a $mse(G_{dr}(\alpha)) = 1$; donc $\frac{x_p}{\bar{x}} = mse(G_{dr}(0)) = 1$ ce qui entraîne $p = 1$, puisque si $p \geq 2$ $x_1 < \bar{x} < x_p$ d'après 2.2.3. (On peut aussi dire que l'on a $m_{dr}(\alpha) = \alpha$ pour tout α dans $[0;1]$ et conclure en utilisant 4.3).
Si $K_p = \frac{x_p}{\bar{x}}$ c'est que $\int_0^1 (mse(G_{pr}(\alpha)) - \frac{x_p}{\bar{x}})d\alpha = 0$, et comme la fonction à intégrer est continue, négative ou nulle, obligatoirement elle est nulle partout et donc $\forall \alpha \in [0;1]$ on a $mse(G_{dr}(\alpha)) = \frac{x_p}{\bar{x}}$; donc $mse(G_{dr}(1)) = \frac{x_p}{\bar{x}}$ ce qui donne encore $\frac{x_p}{\bar{x}} = 1$, et $p = 1$.

5) En posant $I_i = \int_{1-\alpha_{i+1}}^{1-\alpha_i} (\frac{1-a_i-b_i}{\alpha} + a_i)d\alpha$ on a $K_p = \sum_{i=p-1}^{0} I_i$ (voir le 1 ci-dessus).

Puisque $1 - a_{p-1} - b_{p-1} = 0$, $I_{p-1} = \int_0^{1-\alpha_{p-1}} a_{p-1}d\alpha = \frac{n_p x_p}{n\bar{x}}$.

Et pour $i \in \{0;1;\ldots;p-2\}$

$$\begin{aligned} I_i &= (1-a_i-b_i)\ln\frac{1-\alpha_i}{1-\alpha_{i+1}} + a_i(\alpha_{i+1}-\alpha_i) \\ &= (1-a_i-b_i)\ln\frac{1-\alpha_i}{1-\alpha_{i+1}} + \frac{n_{i+1}x_{i+1}}{n\bar{x}} \end{aligned}$$

K_p peut alors s'écrire $\sum_{i=p-2}^{0}(1-a_i-b_i)\ln\frac{1-\alpha_i}{1-\alpha_{i+1}} + \sum_{i=p-1}^{0}\frac{x_{i+1}n_{i+1}}{\bar{x}n}$; mais la dernière somme n'est autre que $\frac{\bar{x}}{\bar{x}} = 1$ et

$$\begin{aligned} 1-a_i-b_i &= 1-a_i-(\beta_i - a_i\alpha_i) \\ &= 1-\beta_i - a_i(1-\alpha_i) \\ &= 1-\beta_i - \frac{x_{i+1}}{\bar{x}}(1-\alpha_i) \end{aligned}$$

ce qui donne finalement $K_p = 1 + \sum_{i=0}^{p-2}(1-\beta_i - \frac{x_{i+1}}{\bar{x}}(1-\alpha_i))\ln\frac{1-\alpha_i}{1-\alpha_{i+1}}$

On peut vérifier que K_p est effectivement supérieur à 1 car pour $0 \le i \le p-2$ on a
$1-\beta_i - \frac{x_{i+1}}{\bar{x}}(1-\alpha_i) = 1-a_i-b_i > 0$ (voir annexe 2) et de $1-\alpha_i > 1-\alpha_{i+1} > 0$ on déduit $\frac{1-\alpha_i}{1-\alpha_{i+1}} > 1$.

Puisque $\alpha_0 = \beta_0 = 0$, le cas particulier $p = 2$ donne :
$K_2 = 1 + (1 - \frac{x_1}{\bar{x}})\ln\frac{1}{1-\alpha_1} = 1 - \frac{\bar{x}-x_1}{\bar{x}}\ln(1-\frac{n_1}{n})$.

6) Si $p = 2$, $x_1 = 0$ et $n_2 = 1$ alors le point précédent permet d'écrire tout de suite :
$K_2 = 1 - \ln(1 - \frac{n-1}{n}) = 1 + \ln n$.

Remarque :
Les groupes ‹‹réels›› constitués de derniers indivdus sont :
le groupe constitué du dernier individu : son *mse* est $\frac{x_2}{\bar{x}} = n$
le groupe constitué des 2 derniers individus : son *mse* est $\frac{\frac{x_2}{2}}{\bar{x}} = \frac{n}{2}$
.....
le groupe constitué des $n-1$ derniers individus : son *mse* est $\frac{n}{n-1}$
le groupe constitué des n derniers individus : son *mse* est 1.
La moyenne des *mse* de ces n groupes est $1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$: elle est différente de $K_2 = 1 + \ln n$. L'explication est que K_2 est la moyenne de tous les $mse(G_{dr}(\alpha))$ pour $\alpha \in [0;1]$, alors que ci-dessus on n'a considéré que les valeurs de α de la forme $\frac{k}{n}$ avec $k \in \{1;2;\ldots;n\}$.
Notons que $\lim_{n\to+\infty} 1 + \ln n - (1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}) = 1 - \gamma$ où γ est la constante d'Euler ($\simeq 0,5772$).

7) On considère toujours l'exemple du 6.1, pour lequel la représentation graphique de la fonction $\alpha \to mse(G_{dr}(\alpha))$ a été faite au 2 ci-dessus.

On a $p = 5$ et $\bar{x} \simeq 26,62$ et donc $K_5 \simeq 1 + \sum_{i=0}^{3}(1-\beta_i - \frac{x_{i+1}}{26,62}(1-\alpha_i))\ln\frac{1-\alpha_i}{1-\alpha_{i+1}}$.
Les 4 termes du $\sum$ sont (se reporter au 2 ci-dessus et au 6.1) :
$(1 - \frac{5}{26,62})\ln\frac{1}{0,8} \simeq 0,1812$
$(0,962 - \frac{15}{26,62}\times 0,8)\ln\frac{0,8}{0,425} \simeq 0,3233$
$(0,752 - \frac{30}{26,62}\times 0,425)\ln\frac{0,425}{0,2} \simeq 0,2058$
$(0,498 - \frac{55}{26,62}\times 0,2)\ln\frac{0,2}{0,075} \simeq 0,0831$
On en déduit $K_5 \simeq 1,79$, valeur bien comprise entre le plus petit *mse* (1) et le plus grand *mse* (3,19).
Remarquons que si l'on fait la moyenne des $mse(G_{dr}(1-\alpha_i))$, pour $i \in \{0;1;2;3;4\}$, c'est-à-dire la moyenne des

ordonnées des points frontières des arcs d'hyperboles formant la représentation graphique de $\alpha \to mse(G_{dr}(\alpha))$ on obtient (voir 7.7) :

$\frac{1+1,2+1,77+2,49+3,19}{5} \simeq 1,93.$

Cette valeur est différente de K_5, toujours pour la même raison : K_5 est la valeur moyenne de tous les $mse(G_{dr}(\alpha))$ pour $\alpha \in [0;1]$.

Pour obtenir une meilleure valeur approchée de K_5 on peut utiliser la méthode des trapézes : on remplace chaque arc d'hyperbole par le segment reliant les deux extrémités de cet arc. On obtient :

$K_5 \simeq 3,19 \times \frac{n_5}{n} + \frac{3,19+2,49}{2} \times \frac{n_4}{n} + \frac{2,49+1,77}{2} \times \frac{n_3}{n} + \frac{1,77+1,2}{2} \times \frac{n_2}{n} + \frac{1,2+1}{2} \times \frac{n_1}{n}$

$K_5 \simeq \frac{3,19 \times 6 + 2,84 \times 10 + 2,13 \times 18 + 1,485 \times 30 + 1,1 \times 16}{80} \simeq 1,83.$

Cette valeur approchée est par excès car les segments considérés plus haut sont au dessus des arcs d'hyperbole.

8) Même démonstration que pour le 3

9) Même démonstration que pour le 4

10) En posant $I_i = \int_{\alpha_i}^{\alpha_{i+1}} \frac{a_i\alpha + b_i}{\alpha} d\alpha$ on a $J_p = \sum_{i=0}^{p-1} I_i$. (voir toujours le 1 ci-dessus)

$I_0 = \int_0^{\alpha_1} a_0 d\alpha = a_0\alpha_1 = \frac{n_1 x_1}{n\bar{x}}$.

Pour $i \in \{1;2;\ldots;p-1\}$ $I_i = a_i(\alpha_{i+1} - \alpha_i) + b_i \ln \frac{\alpha_{i+1}}{\alpha_i} = \frac{n_{i+1}x_{i+1}}{n\bar{x}} + b_i \ln \frac{\alpha_{i+1}}{\alpha_i}$.

D'où $J_p = \frac{n_1 x_1}{n\bar{x}} + \sum_{i=1}^{p-1} \frac{n_{i+1}x_{i+1}}{n\bar{x}} + b_i \ln \frac{\alpha_{i+1}}{\alpha_i} = 1 + \sum_{i=1}^{p-1} (\beta_i - \frac{x_{i+1}}{\bar{x}}\alpha_i) \ln \frac{\alpha_{i+1}}{\alpha_i}$.

On peut vérifier que J_p est bien inférieur à 1 car $\beta_i - \frac{x_{i+1}}{\bar{x}}\alpha_i = b_i < 0$ et $\alpha_{i+1} > \alpha_i$.

Le cas particulier $p = 2$ donne $J_2 = 1 + (\beta_1 - \frac{x_2}{\bar{x}}\alpha_1) \ln \frac{\alpha_2}{\alpha_1}$, et comme $\alpha_1 = \frac{n_1}{n}$, $\alpha_2 = 1$ et $\beta_1 = \frac{n_1 x_1}{n_1 x_1 + n_2 x_2} = \frac{n_1 x_1}{n\bar{x}}$, on obtient $J_2 = 1 - \frac{n_1}{n} \frac{x_1 - x_2}{\bar{x}} \ln \frac{n_1}{n}$.

11) Si $p = 2$, $x_1 = 0$ et $n_2 = 1$ alors, compte tenu de $\bar{x} = \frac{x_2}{n}$, le point précédent permet d'écrire tout de suite :

$J_2 = 1 - \frac{n-1}{n}(-n) \ln \frac{n-1}{n} = 1 + (n-1)\ln(1 - \frac{1}{n})$.

En utilisant le développement limité à l'ordre 2 de la fonction ln on obtient :

$\ln(1 - \frac{1}{n}) = -\frac{1}{n} - \frac{1}{n^2} + \frac{1}{n^2}\varepsilon(n)$ avec $\lim_{n\to+\infty} \varepsilon(n) = 0$

D'où $J_2 = \frac{1}{2n} + \frac{1}{n}(\frac{1}{2n} + \frac{n-1}{n}\varepsilon(n)) = \frac{1}{2n} + \frac{1}{n}\varepsilon_1(n)$ avec $\lim_{n\to+\infty} \varepsilon_1(n) = 0$ et donc lorsque n tend vers $+\infty$ les deux quantités J_2 et $\frac{1}{2n}$ sont équivalentes (c'est-à-dire leur rapport tend vers 1).

Remarque :

Les groupes ‹‹réels›› constitués de premiers indivdus sont :

le groupe constitué du premier individu : son *mse* est 0

le groupe constitué des 2 premiers individus : son *mse* est 0

.....

le groupe constitué des $n-1$ premiers individus : son *mse* est 0

le groupe constitué des n premiers individus : son *mse* est 1.

La moyenne des *mse* de ces n groupes est $\frac{1}{n}$: elle est différente de $J_2 = 1 + (n-1)\ln(1 - \frac{1}{n}) \simeq \frac{1}{2n}$ (pour n grand) toujours parce que J_2 est la moyenne de tous les $mse(G_{pr}(\alpha))$ pour $\alpha \in [0;1]$, alors que ci-dessus on n'a considéré que les valeurs de α de la forme $\frac{k}{n}$ avec $k \in \{1;2;\ldots;n\}$.

ANNEXE 7

mse des séries à valeurs du caractère en progression arithmétique

‖ On considère une série dont les valeurs du caractère sont $x_i = ia + b$ pour $i \in \{1;2;\ldots;p\}$
‖ avec $a > 0$ et $x_1 = a + b > 0$, les effectifs étant tous égaux.
‖ On a alors :

1) ‖ Tout *mse* est inférieur à 2
2) ‖ $\forall \alpha \in]0;1] \lim\limits_{p\to+\infty} mse(G_{dr}(\alpha)) = 2 - \alpha$
‖ en particulier $\lim\limits_{p\to+\infty} c_{50} = 1,5$ et $\lim\limits_{p\to+\infty} c_{10} = 1,9$.
3) ‖ La ***concentration finale*** est toujours faible
‖ mais pour p grand la ***concentration globale*** est toujours moyenne.

Preuve :

1) D'après 2.2.3 le plus fort *mse* est $\frac{x_p}{\bar{x}}$ avec $\bar{x} = \frac{\sum_{i=1}^{p} n_i x_i}{n} = \frac{1}{p}(\sum_{i=1}^{p} x_i) = \frac{x_1 + x_p}{2}$ et donc $\frac{x_p}{\bar{x}} = \frac{2x_p}{x_1 + x_p} < 2$.

2) Si $\alpha = 1$ le résultat est évident puisque $mse(G_{dr}(1)) = 1$ pour tout $p \geq 1$.
On se place dans le cas $\alpha \in]0;1[$.
On a $n = n_1 + n_2 + \ldots + n_p$ avec $n_i = \frac{n}{p}$ pour tout i$\in \{1;2;\ldots;p\}$.
L'effectif (absolu) αn du groupe $G_{dr}(\alpha)$ sera supérieur ou égal au dernier effectif $n_p = \frac{n}{p}$ dès que $\alpha \geq \frac{1}{p}$, ce qui sera réalisé pour p suffisamment grand.
Donc pour p suffisamment grand et compte tenu que $\alpha n < n$, il existe un unique $j \in \{1;2;\ldots;p-1\}$ tel que
$n_{j+1} + n_{j+2} + \ldots + n_p \leq \alpha n < n_j + n_{j+1} + \ldots + n_p$ soit $(p-j)\frac{n}{p} \leq \alpha n < (p-j+1)\frac{n}{p}$
ce qui donne $1 - \alpha \leq \frac{j}{p} < 1 - \alpha + \frac{1}{p}$ et $\lim\limits_{p\to+\infty} \frac{j}{p} = 1 - \alpha$.
Par définition de j on a

$$mse(G_{dr}(\alpha)) = \frac{(\alpha n - (n_{j+1} + n_{j+2} + \ldots + n_p))x_j + \sum_{i=j+1}^{p} n_i x_i}{\alpha \sum_{i=1}^{p} n_i x_i}$$

$$= \frac{(\alpha p - p + j)x_j + \sum_{i=j+1}^{p} x_i}{\alpha \sum_{i=1}^{p} x_i}$$

$$= \frac{((\alpha-1)p + j)(ja + b) + (p-j)\frac{((j+1)a + b + pa + b)}{2}}{\alpha p \frac{a + b + ap + b}{2}}$$

$$= \frac{A + B}{\alpha S}$$

avec $A = ((\alpha - 1)p + j)b + (p - j)(b + \frac{a}{2})$
$B = ((\alpha - 1)p + j)ja + (p - j)\frac{ja + pa}{2} = a(\frac{1}{2}j^2 + \frac{1}{2}p^2 + (\alpha - 1)jp)$
$S = p\frac{a(p+1) + 2b}{2}$
En passant aux limites on a :

$\lim_{p\to+\infty} \frac{A}{p^2} = 0$, $\lim_{p\to+\infty} \frac{B}{p^2} = a(\frac{1}{2}(1-\alpha)^2 + \frac{1}{2} + (\alpha-1)(1-\alpha)) = a(\alpha - \frac{\alpha^2}{2})$

et $\lim_{p\to+\infty} \frac{S}{p^2} = \frac{a}{2}$ et donc $\lim_{p\to+\infty} mse(G_{dr}(\alpha)) = \lim_{p\to+\infty} \frac{A+B}{\alpha S} = 0 + \frac{1}{\alpha} \frac{a(\alpha - \frac{\alpha^2}{2})}{\frac{a}{2}} = 2 - \alpha.$

On en déduit $\lim_{p\to+\infty} mse(G_{dr}(0,5)) = 1,5$ et $\lim_{p\to+\infty} mse(G_{dr}(0,1)) = 1,9$, c'est-à-dire $\lim_{p\to+\infty} c_{50} = 1,5$ et $\lim_{p\to+\infty} c_{10} = 1,9$.

3) D'après le résultat 1, pour tout entier $p \geq 1$ on est sûr que $c_{10} < 2$ et donc la concentration finale est faible, puisque 2 est très inférieur à la valeur maximale 10.
Par contre le fait que c_{50} soit toujours inférieur à 2 ne permet pas d'obtenir une conclusion relative à la concentration globale, puisque sur l'ensemble de toutes les séries la valeur maximale de c_{50} est justemment 2.
Mais pour p grand, d'après le 2 on peut affirmer que $c_{50} \simeq 1,5$ et donc la concentration globale est moyenne.

Remarque :
Il a été vu au 4.6 que $\frac{2}{3}(c_{50} - 1)$ est une valeur approchée du coefficient de Gini g : on en a ici une très bonne illustration, puisque pour la série étudiée ci-dessus il a été prouvé que $\lim_{p\to+\infty} g = \frac{1}{3}$ (voir 6.5), valeur rigoureusement égale à $\frac{2}{3}(\lim_{p\to+\infty} c_{50} - 1)$.

ANNEXE 8

Encadrement du coefficient de Gini en fonction des indicateurs c_{50} et c_{10}

|| Lorsque la répartition n'est pas égalitaire, $g \in [1-t;1-s]$
|| avec $s = \frac{(10b-9)(1,8-b)}{2a+10b-10} - a\frac{0,9a-0,5b}{a-b} + \frac{4,5b-0,1a}{4}$, $t = 0,9a+0,5b+0,1$
|| et $a = m_{pr}(0,5) = 1-0,5c_{50}$, $b = m_{pr}(0,9) = 1-0,1c_{10}$
|| *Remarque 1* : si la répartition est égalitaire, alors $g = 0, c_{50} = 1, c_{10} = 1$ (voir 8.2.1)
|| et dans ce cas le premier terme de s est une forme $\frac{0}{0}$
|| *Remarque 2* : si la répartition tend vers une répartition égalitaire,
|| a et b tendent respectivement vers 0,5 et 0,9, et s et t tendent vers 1,
|| et ainsi on retrouve le fait que g tend vers 0.
|| *Remarque 3* : pour l'exemple du 9.3.2, on obtient comme encadrement $\simeq [0,11;0,2]$
|| alors que $g \simeq 0,15$
|| *Remarque 4* : cet encadrement ne présente pas un gros intérêt car le but de cette étude est de remplacer
|| g par les indicateurs c_{50} et c_{10} ; mais d'un point de vue théorique il confirme que la connaissance
|| de ces indicateurs entraîne pratiquement la connaissance de g (puisque cf le début du 8.2,
|| la connaissance de ces deux indicateurs permet de localiser la courbe de Lorentz).
|| On verra d'ailleurs une petite application de cette annexe au 4) du 9.2.2

Preuve :

On se reportera à la figure du 4.4 où
U est le point d'abscisse 0,5, donc cf 4.2, 9.2.1, son ordonnée est $a = m_{pr}(0,5) = 1-m_{dr}(0,5) = 1-0,5c_{50}$
V est le point d'abscisse 0,9, donc cf 4.2, 9.2.1, son ordonnée est $b = m_{pr}(0,9) = 1-m_{dr}(0,1) = 1-0,1c_{10}$
Cf le 2) de 4.4 la pente de $[OU]$ est $\leq$ à la pente de $[OV]$, cad $\frac{b}{0,9} \geq \frac{a}{0,5}$, ce qui donne $5b \geq 9a$.
On a évidemment $b = m_{pr}(0,9) \geq a = m_{pr}(0,5)$, mais $b = a$ est impossible, car cela conduit à $0 \geq 4a$, soit $a = 0$, ce qui est impossible puisque les effectifs n_i sont non nuls.
Donc $b > a$.
On sait cf 3.1.4 que $g = 2\gamma$ où γ est l'aire de la région située entre le segment $[OA]$ et la courbe de Lorentz (CL).
En notant γ_1 l'aire de la régions située sous (CL) et entre les abscisses 0 et 0,5
γ_2 l'aire de la régions située sous (CL) et entre les abscisses 05 et 0,9
γ_3 l'aire de la régions située sous (CL) et entre les abscisses 0,9et 0,1
on a $\gamma = \frac{1}{2} - \gamma_1 - \gamma_2 - \gamma_3$ soit $g = 1 - 2\gamma_1 - 2\gamma_2 - 2\gamma_3$.
Mais, cf le 3) de 4.4, entre les abscisses 0 et 0,5, (CL) est à l'intérieur du triangleOIU,
entre les abscisses 0,5 et 0,9, (CL) est à l'intérieur du triangle UJV,
entre les abscisses 0,9 et 1, (CL) est à l'intérieur du triangleVKA.
D'où en notant
$U'(0,5,0)$ le projeté orthogonal de U sur $[OB]$, $V'(0,9,0)$ le projeté orthogonal de V sur $[OB]$ et J' le projeté orthogonal de J sur $[OB]$
on a
(1) $aire(IUU') \leq \gamma_1 \leq aire(OUU')$, soit $\frac{(0,5-x_I)\times a}{2} \leq \gamma_1 \leq \frac{0,5a}{2}$
(2) $aire(U'UJJ') + aire(V'VJJ') \leq \gamma_2 \leq aire(U'UVV')$,
soit $\frac{(x_J-0,5)(a+y_J)}{2} + \frac{(0,9-x_J)(y_J+b)}{2} \leq \gamma_2 \leq \frac{(0,9-0,5)(a+b)}{2}$
(3) $aire(V'VKB) \leq \gamma_3 \leq aire(V'VAB)$, soit $\frac{(1-0,9)(b+y_K)}{2} \leq \gamma_3 \leq \frac{(1-0,9)(b+1)}{2}$
Pour préciser ces différentes encadrements, il nous faut les coordonnées de I, J, K.

L'équation de la droite (UV) étant $(x-\frac{1}{2})(b-a)-(y-a)(\frac{9}{10}-\frac{1}{2})=0$,

de $y_I = o$ on tire $x_I = \frac{0,9a-0,5b}{a-b}$ (licite car on a vu plus haut que $b > a$),

de $x_K = 0$ on tire $y_K = \frac{5b-a}{4}$.

Reste à trouver les coordonnées de J qui est le point d'intersection des droites (OU) et (AV), lesquelles ont pour équations respectives $y = 2ax$ et $y = 10(1-b)x + 10b - 9$.

D'où $x_J = \frac{10b-9}{2a+10b-10}$ et $y_J = \frac{2a(10b-9)}{2a+10b-10}$, calcul qui sera licite si $2a+10b-10 \neq 0$.

$2a+10b-10 = \frac{m_{pr}(0,5)}{0,5} + 9\frac{m_{pr}(0,9)}{0,9} - 10 \leq 1+9-10 = 0$, puisque cf 4.3, $m_{pr}(\alpha) \leq \alpha$.

Et $2a+10b-10$ ne sera nul que si $m_{pr}(0,5) = 0,5$ et $m_{pr}(0,9) = 0,9$, cad , cf 4.3, si la répartition est égalitaire, ce qui a été exclu par hypothèse.

Donc $2a+10b-10 < 0$ et les coordonnées de J ci-dessus sont bien licites.

On peut maintenant encadrer $1-g = 2\gamma_1 + 2\gamma_2 + 2\gamma_3 : 1-g \in [s;t]$, cad $g \in [1-t;1-s]$, avec

$s = (0,5-x_I)a + (x_J-0,5)(a+y_J) + (0,9-x_J)(y_J+b) + 0,1(b+y_K)$

$t = 0,5a + 0,4(a+b) + 0,1(b+1) = 0,9a + 0,5b + 0,1$

Reste à écrire s uniquement en fonction de a et b :

$s = (a-b)x_J + 0,4y_J - ax_I + 0,1y_K + b$

$s = \frac{(a-b)(10b-9) + 0,4\times 2a(10b-9)}{2a+10b-10} - a\frac{0,9a-0,5b}{a-b} + 0,1\frac{5b-a}{4} + b$

$s = \frac{(10b-9)(1,8a-b)}{2a+10b-10} - a\frac{0,9a-0,5b}{a-b} + \frac{4,5b-0,1a}{4}$

On a bien obtenu l'encadrement annoncé.

Etudions maintenant le comportement des bornes de cet encadrement lorsque la répartition devient égalitaire.

Dans ce cas (CL) devient très proche du segment diagonal $[OA]$ et pour tout α dans $[0;1]$, $m_{pr}(\alpha) \simeq \alpha$, donc a et b tendent vers respectivement vers 0,5 et 0,9.

Donc t tend vers 0,45+0,45+0,1=1.

Quant à s, le premier terme devient une forme $\frac{0\times 0}{0}$; en fait cf 4.3, puisque la répartition n'est pas égalitaire, $a = m_{pr}(0,5) < 0,5$ et $b = m_{pr}(0,9) < 0,9$

et ce premier terme peut s'écrire $\frac{1,8a-b}{1+\frac{2a-1}{10b-9}}$ avec $\frac{2a-1}{10b-9} > 0$.

D'où $|\frac{1,8a-b}{1+\frac{2a-1}{10b-9}}| < |1,8a-b|$ qui tend vers 0 lorsque la répartition devient égalitaire et alors s tend vers

$0 - 0,5\frac{0,45-0,45}{-0,4} + \frac{4,05-0,05}{4} = 1.$

Cas de l'exemple 9.3.2 :

on a $c_{10}1,39$, $c_{50} = 1,2$ donc $a = 0,4$, $b = 0,861$

$t = 0,9\times 0,4 + 0,5\times 0,861 + 0,1 = 0,8905$ et $1-t \simeq 0,11$

$s = \frac{0,05499}{-0,59} - 0,4\frac{-0,0705}{-0,461} + \frac{3,8345}{4} = 0,80425$ et $1-t \simeq 0,2$

Ce qui donne pour g l'encadrement $\simeq [0,11;0,2]$.

BIBLIOGRAPHIE

[1] D'AGOSTINO S., LAUREYS G., SIMON E., TROMBERT G.
Les épreuves de sciences économiques et sociales au bac ES.
Des méthodes pour réussir.
Ed : Vuibert 1995

[2] ANTIBI A., BARRA R., MALAVAL J.
Math 1ére ES.
Ed : Nathan 1993

[3] COMBROUZE A.
Probabilités et statistiques.
Ed : Puf 1993

[4] DEHEUVELS P.
L'intégrale.
Ed : Puf 1980

[5] MAZIERI W.
Notions essentielles de statistiques et calcul des probabilités.
Ed : Sirey 1967

[6] PY B.
Statistiques descriptives. Nouvelle méthode pour bien comprendre et réussir.
Ed : Economica 1990