Sommaire Site A P

Présentation du problème

(avant d'en venir à la méthode *mse*)

On considère une série d'individus dont on étudie un caractère quantitatif prenant les valeurs $x_1,x_2,...,x_p$ avec $x_1 >0$, $x_i < x_{i+1}$ et on s'intéresse au problème 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 Gini-Lorenz qui consiste à faire une courbe (la courbe de Lorenz) et/ou à calculer le coefficient de Gini, noté *g*.

Note : *(cf fr.wikipedia.org)*
Max Otto Lorenz (1880-1962), américain, proposa sa courbe en 1905 (sur le net, on rencontre beaucoup plus souvent courbe de Lorenz que courbe de Lorentz) et,
Corrado Gini (1884-1965), italien, a introduit son coefficient en 1914.

Au préalable pour ceux ne connaissant pas la méthode de Gini-Lorenz voici un petit exposé : coefficient de Gini ; cependant il n'est pas indispensable à lire, puisque la méthode *mse* est indépendante de la méthode Gini.

On a toujours $0 \leq g < 1$, et donc, contrairement à ce qui est parfois dit il ne peut jamais prendre la valeur 1 ; il est toujours inférieur à 0,9 pour les séries décilées et de façon générale, j'ai prouvé que :

$$g \leq \frac{\sqrt{x_p} - \sqrt{x_1}}{\sqrt{x_p} + \sqrt{x_1}} < 1$$

(formule que je n'ai lue nulle part, mais je n'ai pas la prétention d'avoir tout lu sur Gini)

Le coefficient de Gini de part sa nature intégrale est une valeur moyenne : au prix de calculs compliqués, (ce qui est un 1er inconvénient), on perd de l'information.

On peut le considérer comme un indicateur de concentration global

En fait il est pratiquement égal, à une transformation affine près, à la masse possédée par les 50% derniers individus : pourquoi ne pas considérer alors directement ce chiffre? Cela aurait le mérite de simplifier les calculs!

Mais pour autant, on ne ferait pas disparaître le 2ième inconvénient de ce coefficient : à savoir qu'il n'apporte pas d'éclairage précis sur la série.

En effet, lorsqu'on a trouvé g=0,66 (série des patrimoines de 1986) 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.

Notons que le coefficient e=l'écart maximum entre masse et effectif est considérablement plus simple à calculer que le coefficient de Gini et s'analyse exactement de la même façon : si e=0 il y a répartition égalitaire et si e est proche de 1 beaucoup ont peu ou peu ont beaucoup. Compte tenu de sa simplicité il peut remplacer avantageusement le coefficient de Gini .

Cependant ces deux coefficients ne permettent pas une analyse précise de la série (en terme de concentration), cela parce que ils reposent tous les deux sur la même idée : pour comparer deux nombres on fait leur différence.

Or si l'on considère 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 (60/30)?

En général c'est plutôt des rapports que l'on considère : par exemple le rendement d'un appareil électrique 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 bénéfice net par action (rapport qui peut être beaucoup plus grand que 1) et non la différence de ces deux 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*)

Nouvelle méthode d'analyse de la concentration d'une série :

la méthode *mse*

J'ai terminé la mise au point de cette méthode en 1998, la mise en ligne de ce **résumé** datant de janvier 2000.

L'exposé complet, en .pdf, de la méthode mse, avec toutes les justifications et davantage d'exemples a été mis en ligne en novembre 2012 : voir ici

Plan

A : Notations, définitions, propriétés (les démonstrations ne sont pas ici) : il s'agit de justifier brièvement la méthode.

B : Mise en oeuvre pratique : calculs à effectuer, analyse des résultats.

C : Exemples

D : Conclusion

A Notations, définitions, propriétés (sans démo)

1) La série est constituée de n individus prenant les p valeurs du caractère $x_1,x_2,...,x_p$ avec comme effectifs correspondants $n_1,n_2,...,n_p$.

On suppose $1 \leq p$, $x_1 > 0$, $x_i < x_{i+1}$ et tous les effectifs sont strictement positifs.

Le cas $x_1 = 0$ sera parfois envisagé mais cela sera alors explicitement dit.

On a évidemment $n = n_1+n_2+...+n_p$.

2) La masse totale de la série est notée $m = n_1x_1+...+n_px_p$.

3) La moyenne de la série est notée moy = m/n.

4) Soit G un sous-groupe d'individus de la série : le rapport égal à la masse (en pourcentage de la masse totale) possédée par G divisée par l'effectif (en pourcentage de l'effectif total) du groupe G est noté *mse(*G).

Par exemple si les 20% derniers individus (cad les 20% des n individus ayant les plus fortes valeurs du caractère) possèdent 50% de la masse, leur *mse* est 2,5 : ils ont 2,5 fois plus en masse qu'en effectif. Cela est tout de même plus parlant que de dire : le groupe des 20% derniers individus possédé une masse supérieure de 0,3 à son effectif. C'est cet aspect qui justifie (à mon avis) le fait de considérer ces rapports masse sur effectif au lieu de considérer les différences masse moins effectif (méthode Gini-Lorenz).

5) *mse(*G)=moyenne du groupe G divisé par moy

*mse(*G) est toujours dans l'intervalle [x_1/moy ; x_p/moy]

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

6) 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* égal à 1 : dans ce cas il n'y a pas de phénomène de concentration.

Cette situation a lieu si et seulement si tous les individus ont la même valeur du caractère (p=1).

7) L'objectif est de savoir s'il existe des groupes ayant beaucoup plus en masse qu'en effectif (voir présentation du problème), c'est-à-dire s'il existe des groupes ayant des *mse* élevés. Or ce sont les groupes constitués des derniers individus qui ont les *mse* les plus élevés : de façon plus précise si G est un groupe d'effectif (en pourcentage) supérieur ou égal à α (α est un nombre compris entre 0 et 1) alors le *mse* du groupe constitués des α derniers individus est supérieur ou égal à *mse(*G).

Pour cette raison on va s'intéresser uniquement aux *mse* des groupes constitués des derniers individus : on notera *mse*(G_dr(α)) le *mse* du groupe constitué des α

derniers individus (les individus étant classés par valeur croissante du caractère et α représentant un effectif, non nul, en pourcentage).

Notons que, à α fixé, si *mse*(G_dr(α)) augmente cela veut dire que la masse du groupe G_dr(α) augmente et donc se concentre (au sens habituel du terme) sur ce groupe.

On a toujours $1 \leq$ *mse*(G_dr(α))

Et un aspect essentiel :

si α est dans]0;1[alors *mse*(G_dr(α)) = 1 si et seulement si la répartition est égalitaire (p=1)

On peut donc considérer que *mse*(G_dr(α)) est un indicateur de concentration du groupe G_dr(α) et à ce titre, connaître l'état de concentration c'est connaître à priori tous ces *mse*

Cependant on peut se contenter de calculer le *mse* des 50% derniers individus et celui des 10% derniers individus.

3 raisons à cela :

==> toutes les séries ayant respectivement ces mêmes *mse* auront des courbes de Lorenz 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 répartition de la masse (la courbe de Lorenz est caractéristique de la série à deux coefficients de proportionnalité près).

==> très souvent pour insister sur le fait qu'une série est très concentrée on cite la masse possédée par les 10% derniers individus.

==>le coefficient de Gini est approximativement égal à 2(*mse*(G_dr(0,5))-1)/3 : c'est-à-dire le coefficient de Gini (pas facile à calculer) est pratiquement égal, à une transformation affine près, au *mse* des 50% derniers individus (très facile à calculer)!

On représentera donc la concentration d'une série par un vecteur C à 2 composantes : le *mse* des 50% derniers individus et le *mse* des 10% derniers individus.

On peut considérer que l'information "globale" donnée par le *mse* des 50% derniers individus 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 moyenne.

Bien entendu des séries peuvent avoir le même *mse*(G_dr(0,5)) (et donc elles auront à peu près le même coefficient de Gini) sans pour autant avoir les mêmes *mse*(G_dr(0,1)), ce qui permet une meilleure différentiation de ces séries.

B Mise en oeuvre pratique de la méthode *mse*

Phase 1 Calcul du vecteur concentration $C(c_{50},c_{10})$

c_{50} = *mse*(G_dr(0,5)) = 2 fois la masse (en pourcentage) possédée par les 50% derniers individus

c'est l'indicateur de concentration des 50% derniers individus.

c_{10} = *mse*(G_dr(0,1)) = 10 fois la masse (en pourcentage) possédée par les 10%derniers individus

c'est l'indicateur de concentration des 10% derniers individus.

Notons tout de suite que pour une série décilée ces deux *mse* peuvent se calculer de tête : la première composante est le double de la somme des masses (en pourcentage) des 5 dernières classes, la deuxième composante étant la masse (en pourcentage) de la dernière classe. Pour une série quelconque on verra sur les exemples que le calcul n'est pas bien long : une ligne pour chaque composante.

Phase 2 Analyse des résultats

Cette analyse repose sur les diverses propriétés de ces deux composantes :

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

Sur l'ensemble de toutes les séries c_{50} décrit tout l'intervalle [1;2[

Sur l'ensemble de toutes les séries c_{10} décrit tout l'intervalle [1;10[

La répartition est égalitaire si et seulement si $c_{50}=1$ ou si et seulement si $c_{10}=1$

Si $x_1=0$ alors c_{50} peut prendre la valeur 2 si et seulement si les 50% premiers individus ont 0

Si $x_1=0$ alors c_{10} peut prendre la valeur 10 si et seulement si les 90% premiers individus ont 0

On peut donner des qualificatifs aux diverses situations, par exemple :

Qualificatifs de concentration globale

$c_{50} = 1$	répartition égalitaire ou concentration nulle
$c_{50} \cong 1,5$	concentration globale moyenne
$c_{50} \cong 2$	concentration globale maximum

Qualificatifs de concentration finale

$c_{10} = 1$	répartition égalitaire ou concentration nulle

$c_{10} \cong 5{,}5$ concentration finale moyenne

$c_{10} \cong 10$ concentration finale maximum

Notons aussi que c_{10} est le plus grand *mse* parmi tous les groupes d'effectifs supérieur ou égal à 10% et, 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.

Remarque :

Complétons les résultats ci-dessus par :

On a toujours $1 \leq c_{50} \leq c_{10} \leq 9 \times c_{50} - 8 < 10$ avec

$c_{50} = c_{10}$ si et seulement si la répartition est égalitaire à l'intérieur du groupe des 50% derniers individus

$c_{10} = 9 \times c_{50} - 8$ si et seulement si la répartition est égalitaire à l'intérieur du groupe des 90% premiers individus

C Exemples

Exemple 1

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

classes	x_i	n_i	$n_i x_i$	$\Sigma-$	$\Sigma-$
[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		

L'avant dernière colonne est constituée des effectifs cumulés décroissants.

La dernière colonne est constituée des masses cumulées décroissantes.

Calcul des deux composantes du vecteur concentration :

50%n = 88 et puisque 37 < 88 <176 on immédiatement :

c_{50} = 2×((1521000+(88-37)×7500)/2563500) soit environ 1,58

(les 50 % derniers individus ont 1,58 fois plus en masse qu'en effectif)

10%n = 17,6 et puisque 9 < 17,6 <25 on a immédiatement

c_{10} = 10×((651000+(17,6-9)×37500)/2563500) soit environ 3,8

(les 10% derniers individus ont 3,8 fois plus en masse qu'en effectif)

d'où :

la concentration globale est un peu au dessus de la moyenne (1,58 > 1,5)

la concentration finale est en dessous de la moyenne (3,8 < 5,5)

Et évidemment le plus fort *mse* parmi tous les groupes d'effectifs supérieur ou égal à 10% est 3,8

Remarque 1 :

Le plus fort *mse* de la série n'est pas ici 3,8 puisque l'effectif des n_p = 2 derniers individus (1,1% de l'effectif total) n'est pas inférieur à 10%. Le plus fort *mse,* parmi tous les groupes de la série, est justement le *mse* des 2 derniers individus égal à x_p/moy soit environ 6,7 : ils ont 6,7 fois plus en masse qu'en effectif. mais ceci est à relativiser car la plage de variation du *mse* des 1,1% derniers individus est [1;1/1,1%[soit environ [1;91[, et donc la concentration au niveau du groupe des 1,1% derniers individus est très faible.

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

Remarque 2 :

Pour ceux connaissant la méthode de Gini-Lorenz, je leur laisse le soin de vérifier que le coefficient de Gini est environ égal à 0,41 : après beaucoup de calculs (en tout cas beaucoup plus que ceux nécessaires au calcul de c_{50} et c_{10}) on ne peut que conclure à "concentration" en dessous de la moyenne (puisque 0,41 < 0,5), cela sans rien savoir de précis sur les groupes ayant beaucoup plus en masse qu'en effectif!

Enfin on peut vérifier ici l'approximation annoncée plus haut : une valeur approchée du coefficient de Gini est 2×(c_{50}-1)/3.

En effet pour cet exemple 2(c_{50}-1)/3 = 2×0,58/3 soit environ 0,39, le coefficient de Gini étant 0,41.

Soyons honnête : cette approximation n'est pas toujours aussi bonne. Mais son intérêt est de montrer que calculer le coefficient de Gini c'est pratiquement calculer (par un chemin détourné et lourd) la masse des 50% derniers individus.

Exemple 2

Il s'agit des séries des revenus et patrimoines 1986, présentées sous formes 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 on obtient :

pour les revenus

c_{50} = *mse*(G_dr(0,5)) = 2×((27,8+16,1+12,6+10,6+8,8)/100) ≅ 1,52

c_{10} = *mse*(G_dr(0,1)) = 10×27,8/100 ≅ 2,78

la concentration globale est moyenne et la concentration finale est bien en dessous de la moyenne

pour les patrimoines

c_{50} = *mse*(G_dr(0,5)) = 2×((53,8+15,1+10,6+8,6+5,9)/100) ≅ 1,88

c_{10} = *mse*(G_dr(0,1)) = 10×53,8/100 ≅ 5,38

la concentration globale est extrême (on n'est pas loin de la valeur maximum) 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} représente le plus fort *mse* pour chaque série : par exemple pour la série des patrimoines le pus grand rapport masse sur effectif pour tous les groupes possibles est 5,38.

Remarque 2

Avec beaucoup plus de calculs la méthode de Gini conduit à une conclusion moins nuancée : la série des patrimoines est beaucoup plus "concentrée" que la série des revenus puisque $g_{revenus}$ = 0,37 alors que $g_{patrimoines}$ = 0,66. En outre les chiffres 0,37 et 0,66 n'ont pas d'interprétation économique précise contrairement à c_{50} et c_{10} qui sont des *mse.*

Remarque 3

Il existe des séries plus concentrées que la série des revenus au niveau des 10% derniers, par exemple en 1999 "65% de l'encours global des actions, obligations et autre OPCVM se trouve

logé dans quelques 11% seulement des comptes titres ouverts dans les banques et sociétés de Bourse françaises" (d'après un journal économique d'août 1999). On a donc ici *mse*(G_dr(0,11)) = 65/11 et donc c_{10} = *mse*(G_dr(0,10))>=65/11 ≅ 5,91.

Exemple 3

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

x_i est le revenu annuel net en milliers de francs

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

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

$n_i x_i$ est exprimé en milliards de francs.

x_i	n_i	$n_i x_i$	Σ–	Σ–
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		

Les deux dernières colonnes sont toujours les effectifs et masses en cumulés décroissants.

Calcul des deux composantes du vecteur concentration :

50%n = 18,6 d'où c_{50} = 2×((2857+(18,6-14,3)×84)/4379) ≅ 1,47

10%n = 3,72 d'où c_{10} = 10×((873+(3,72-1,75)×216)/4379) ≅ 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 la concentration finale a un peu augmenté (c_{10} passe de 2,78 à 3).

Cela signifie que les 50% derniers ont toujours à peu près pareil mais les 10% derniers ont un peu plus, cela au détriment des 80% premiers des 50% derniers!

Remarque

La comparaison des coefficients de Gini ne permettrait évidemment pas une telle conclusion.

Exemple 4

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_ix_i est exprimé en millions de francs.

x_i	n_i	n_ix_i	$\Sigma-$	$\Sigma-$
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		

Les deux dernières colonnes sont toujours les effectifs et masse en cumulés décroissant.

Calcul des deux composantes du vecteur concentration :

50%n= 2053,5 d'où $c_{50}=2\times((1130040+(2053,5-730)\times700)/2826871)\cong1,46$

10%n=410,7 d'où $c_{10}=10\times((410,7\times1548)/2826871)\cong2,25$

Si on compare avec les revenus français 1986 (voir exemple 2) on a une même concentration globale : elle est moyenne, c_{50} étant à peu près égal à 1,5 dans les 2 cas ; quant à la concentration finale, faible dans les 2 cas , elle est tout de même plus faible pour les belges ($c_{10}\cong2,25$) que pour les français ($c_{10}\cong2,78$).

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 $2(c_{50}-1)/3=2\times0,46/3$ soit environ 0,31 : on retrouve l'ordre de grandeur du coefficient de Gini et donc, comme déjà dit en remarque 2 de l'exemple 1, calculer le coefficient de Gini c'est pratiquement calculer la masse des 50% derniers individus, cela par un chemin détourné et lourd.

D Conclusion

Cette méthode *mse* me paraît beaucoup plus simple à mettre en oeuvre que celle du coefficient de Gini et elle permet des conclusions beaucoup plus précises.

Mais évidemment je suis un peu (?) partial, puisque c'est "ma" méthode ; c'est donc maintenant au lecteur de juger!

Exposé complet, en .pdf, dela méthode *mse*, avec toutes les justifications et davantage d'exemples