REVUE DE STATISTIQUE APPLIQUÉE

J. L. PETIT

Généralisation du paradoxe de Simpson

Revue de statistique appliquée, tome 40, nº 3 (1992), p. 47-61
<http://www.numdam.org/item?id=RSA_1992__40_3_47_0>

Rev. Statistique Appliquée, 1992, **XXXX** (3), 47-61

GÉNÉRALISATION DU PARADOXE DE SIMPSON

J.L. Petit

Centre de Recherche en Applications Statistiques
Ecole Nationale Supérieure Polytechnique
B.P. 8390 Yaoundé – Cameroun

RÉSUMÉ

Le Paradoxe de Simpson est produit par la notion de probabilité conditionnelle. Nous présentons ce phénomène à partir d'un exemple et nous en donnons quelques propriétés. Nous généralisons ce paradoxe en utilisant la notion de croissance stochastique.

Mots-clés : *Paradoxe de Simpson, Probabilité Conditionnelle, Croissance Stochastique.*

ABSTRACT

The Simpson's Paradox is produced by the notion of conditional probability. We introduce this paradox by an exemple and give some properties. We generalise this paradox with the notion of stochastic order.

Key-words : *Simpson's Paradox, Conditional Probability, Stochastic Order.*

Introduction

Nous allons étudier et généraliser le classique Paradoxe de Simpson (cf : [1],[4],[6]). Ce paradoxe est produit par la notion de probabilité conditionnelle et consiste en le phénomène suivant :

Soit une population Ω sur laquelle on mesure trois caractères X, Y, Z. On veut étudier le caractère X par rapport au caractère Y. Pour effectuer une telle étude, il se présente deux méthodes :

Etude Marginaliste : on étudie la distribution conditionnelle de X par rapport à Y $(P(X/Y))$.

Etude non-Marginaliste : on étudie la distribution conditionnelle de X par rapport à Y et Z $(P(X/Y, Z))$.

Le Paradoxe de Simpson consiste en l'effet suivant : les deux études conduisent à des conclusions opposées.

1. Exemple de paradoxe de Simpson

Nous allons étudier une population Ω de Lycéens dans la ville d'Anchourie passant un certain examen.

Remarques :

Le nom d'Anchourie revient souvent dans le livre d'Engel (cf : [2]).

On cherche à comparer les résultats des filles aux résultats des garçons. Sur la population Ω on mesure trois caractères :

$$L : \Omega \to \{b, c\} : \text{Lycée} ; b = \text{Buffon}, c = \text{Condorcet}$$
$$R : \Omega \to \{d, e\} : \text{Réussite} ; d = \text{diplôme}, e = \text{échec}$$
$$S : \Omega \to \{f, g\} : \text{Sexe} ; f = \text{fille}, g = \text{garçon}.$$

Nous utiliserons les notions suivantes :

$$N(r, s, l) = \left| R^{-1}(r)(S^{-1}(s)(L^{-1}(l) \right|$$

Il existe deux méthodes classiques pour comparer deux nombres réels x et y :

la comparaison additive : comparer $x - y$ à 0

la comparaison multiplicative : comparer x/y à 1.

Remarques :

La comparaison additive est une notion affine : invariance par translation. La comparaison multiplicative est une notion projective : invariance par homothétie.

1.1 Etude de R en fonction de S

On élimine le caractère L par sommation :

$$N(r, s, \cdot) = \sum_{1} N(r, s, l)$$

Le Tableau (1) donne les effectifs $N(r, s, \cdot)$ pour la ville d'Anchourie :

Tableau 1
Effectifs $N(r, s, \cdot)$ pour la ville d'Anchourie

R \ S	f	g	t
d	46	20	66
e	74	40	114
t	120	60	180

Comparaison additive :

Pour le diplôme : $20 < 46 \Longleftrightarrow N(d, g, \cdot) \underset{+,d}{<} N(d, f, \cdot)$

Pour l'échec : $40 < 74 \Longleftrightarrow N(e, g, \cdot) \underset{+,e}{<} N(e, f, \cdot)$

Remarques :

Si nous avons la condition de normalisation :

$$N(\cdot, g, \cdot) = N(\cdot, f, \cdot)$$

alors les deux ordres $\underset{+,d}{<}$ et $\underset{+,e}{<}$ sont duaux l'un de l'autre.

Comparaison multiplicative :

Pour comparer les résultats en fonction du sexe, nous allons prendre les distributions conditionnelles $P(R/S)$ (cf Tableau (2)) :

$$P(r/s) = \frac{N(r, s, \cdot)}{N(\cdot, s, \cdot)}$$

Tableau 2

Probabilités Conditionnelles $P(r/s)$ pour la ville d'Anchourie.

R \ $P(\ /S)$	f	g
d	23/60	1/3
e	37/60	2/3
t	1	1

Nous obtenons ainsi :

Pour le diplôme : $1/3 < 23/60 \Longleftrightarrow P(d/g) \underset{*,d}{<} P(d/f)$

Pour l'échec : $37/60 < 2/3 \Longleftrightarrow P(e/f) \underset{*,e}{<} P(e/g)$

Remarques :

A cause de la normalisation des probabilités, les deux ordres $\underset{*,d}{<}$ et $\underset{*,e}{<}$ sont en dualité. Les deux ordres $\underset{+,d}{<}$ et $\underset{*,d}{<}$ sont identiques si on a la condition de normalisation précédente.

Conclusion :

Pour la comparaison multiplicative, nous avons le résultat suivant :

Dans la ville d'Anchourie les filles réussissent mieux que les garçons.

Remarques :

La comparaison additive ne permet pas d'apporter une conclusion : les filles obtiennent plus de diplômes et d'échecs que les garçons.

1.2 Etude de R en fonction de S et de L

Les tableaux (3) et (4) donnent les effectifs $N(r,s,b)$ et $N(r,s,c)$ correspondant aux deux lycées b et c :

Tableau 3
Effectifs $N(r,s,b)$ pour le lycée Buffon

R \ S	f	g	t
d	1	5	6
e	11	25	36
t	12	30	42

Tableau 4
Effectifs $N(r,s,c)$ pour le lycée Condorcet

R \ S	f	g	t
d	45	15	60
e	63	15	78
t	108	30	138

A partir des tableaux (1), (3) et (4) on vérifie la relation :

$$N(r,s,b) + N(r,s,c) = N(r,s,.)$$

Nous n'allons pas traiter la comparaison additive mais seulement la comparaison multiplicative.

Comparaison multiplicative :

Les tableaux (5) et (6) donnent les distributions conditionnelles $P(R/S,L)$:

$$P(r/s,l) = \frac{N(r,s,l)}{N(\cdot,s,l)}$$

Tableau 5
Probabilités conditionnelles $P(r/s,b)$ pour le lycée Buffon

R \ $P(\ /s,b)$	f	g
d	1/12	1/6
e	11/12	5/6
t	1	1

Tableau 6
Probabilités conditionnelles $P(r/s,c)$ pour le lycée Condorcet

R \ $P(\ /s,c)$	f	g
d	5/12	1/2
e	7/12	1/2
t	1	1

Nous obtenons ainsi :

$$1/12 < 1/6 \Longleftrightarrow P(d/f,b) \underset{*,d}{<} P(d/g,b)$$
$$5/12 < 1/2 \Longleftrightarrow P(d/f,c) \underset{*,d}{<} P(d/g,c)$$

Conclusion :

Pour la comparaison multiplicative, nous obtenons un exemple de Paradoxe de Simpson :

Dans la ville d'Anchourie, les filles réussissent mieux que les garçons.

Dans les lycées Buffon et Condorcet, les garçons réussissent mieux que les filles.

Le Paradoxe de Simpson se traduit par le système d'inégalités :

$$\frac{N(d,f,b)}{N(\cdot,f,b)} < \frac{N(d,g,b)}{N,\cdot,g,b)}$$
$$\frac{N(d,f,c)}{N(\cdot f,c)} < \frac{N(d,g,c)}{N(\cdot,g,c)}$$
$$\frac{N(d,f,\cdot)}{N(\cdot,f,\cdot)} > \frac{N(d,g,\cdot)}{N(\cdot,g,\cdot)}$$

Remarques :

Le Paradoxe de Simpson ne peut pas être produit par la comparaison additive :

$N(d,f,b) < N(d,g,b)$ et $N(d,f,c) < N(d,g,c)$ entraînent $N(d,f,.) < N(d,g,.)$

2. Opération médiane

Introduisons une opération importante sur le corps des rationnels $\mathbb{Q}$, l'opération médiane, notée $\square$, et définie par :

$$\frac{n_1}{m_1} \square \frac{n_2}{m_2} = \frac{n_1+n_2}{m_1+m_2}$$

Remarques :

1) Cette opération n'est pas définie sur $\mathbb{Q}$ mais sur $\mathbb{N}^2$. Pour la définir sur $\mathbb{Q}$, il faut considérer des fractions irréductibles.

2) Cette opération est fondamentale dans l'étude des *suites de Farey* (cf : [3]).

3) Cette opération se rencontre en statistique. Considérons, par exemple, un ensemble de pays et pour chaque pays définissons les deux caractères suivants :

$$S(\omega) = \text{Surface de } \omega$$
$$H(\omega) = \text{Nombre d'habitants de } \omega$$

Alors le caractère H/S vérifie :

$$H/S(\omega_1 \cup \omega_2) = H/S(\omega_1) \square H/S(\omega_2)$$

Donnons quelques propriétés de cette opération médiane :

Commutative.

Associative.

Propriété barycentrique :

$$\frac{n_1}{m_1} \square \frac{n_2}{m_2} = \left(\frac{m_1}{m_1+m_2}\right)\frac{n_1}{m_1} + \left(\frac{m_2}{m_1+m_2}\right)\frac{n_2}{m_2}$$

Relations avec la structure d'ordre :

$$\min\left(\frac{n_1}{m_1}, \frac{n_2}{m_2}\right) \leqslant \frac{n_1}{m_1} \square \frac{n_2}{m_2} \leqslant \max\left(\frac{n_1}{m_2}, \frac{n_2}{m_2}\right)$$

$$\frac{n_1}{m_1} \leqslant \frac{n_1'}{m_1'},\ \frac{n_2}{m_2} \leqslant \frac{n_2'}{m_2'} \text{ n'entraîne pas } \frac{n_1}{m_1} \square \frac{n_2}{m_2} \leqslant \frac{n_1'}{m_1'} \square \frac{n_2'}{m_2'}$$

Remarques :

C'est cette non compatibilité entre l'ordre sur $\mathbb{R}$ et l'opération $\square$ qui produit le Paradoxe de Simpson.

3. Etude du paradoxe de Simpson

Sur une population Ω on étudie 3 caractères binaires :

$$X : \Omega \to \{x_1, x_2\}$$
$$Y : \Omega \to \{y_1, y_2\}$$
$$Z : \Omega \to \{z_1, z_2\}$$

Cette étude nous fournit les effectifs :

$$N(x_i, y_j, z_k)$$

Le Paradoxe de Simpson se traduit par les inégalités suivantes :

$$\frac{N(x_1,y_1,z_1)}{N(\cdot,y_1,z_1)} = \alpha < \frac{N(x_1,y_2,z_1)}{N(\cdot,y_2,z_1)} = \gamma$$

$$\frac{N(x_1,y_1,z_2)}{N(\cdot,y_1,z_2)} = \beta < \frac{N(x_1,y_2,z_2)}{N(\cdot,y_2,z_2)} = \delta$$

$$\frac{N(x_1,y_1,\cdot)}{N(\cdot,y_1,\cdot)} = \frac{N(x_1,y_1,z_1)}{N(\cdot,y_1,z_1)} \square \frac{N(x_1,y_1,z_2)}{N(\cdot,y_1,z_2)} > \frac{N(x_1,y_2,\cdot)}{N(\cdot,y_2,\cdot)}$$

$$= \frac{N(x_1,y_2,z_1)}{N(\cdot,y_2,z_1)} \square \frac{N(x_1,y_2,z_2)}{N(\cdot,y_2,z_2)}$$

En posant :

$$\lambda = \frac{N(\cdot, y_1, z_1)}{N(\cdot, y_1, \cdot)}; \ \mu = \frac{N(\cdot, y_2, z_1)}{N(\cdot, y_2, \cdot)}$$

le Paradoxe de Simpson s'écrit donc :

$$\alpha < \gamma, \beta < \delta, \lambda\alpha + (1 - \lambda)\beta > \mu\gamma + (1 - \mu)\delta$$

où les 6 nombres $\alpha, \beta, \gamma, \delta, \lambda, \mu$ appartiennent à $[0, 1]$.

Ce paradoxe ne peut donc se produire que dans les deux cas suivants :

1er cas : $0 < \alpha < \gamma < \beta < \delta < 1$

2e cas : $0 < \beta < \delta < \alpha < \gamma < 1$

On passe d'un cas à l'autre par les transpositions :

$$\alpha \Longleftrightarrow \beta; \ \gamma \Longleftrightarrow \delta$$

La zone hachurée de la figure (1) représente les points où se produit le Paradoxe de Simpson, quand on fixe les paramètres α, β, γ, δ, dans le 1er cas :

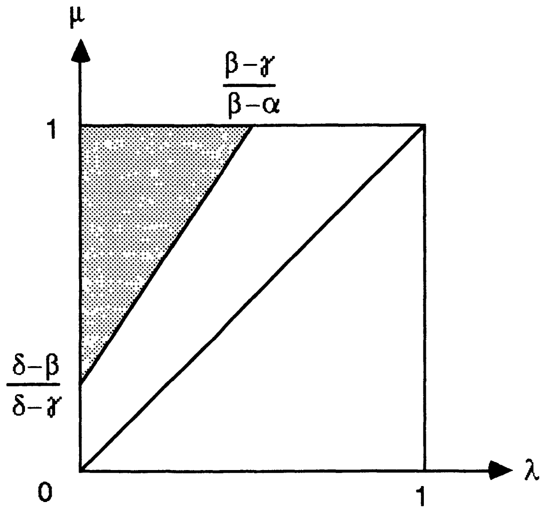

Figure 1
Paradoxe de Simpson

Remarque :

Si on fixe α, β, γ, δ et si on suppose une distribution uniforme pour le couple (λ, μ), on peut dire que la probabilité d'apparition du Paradoxe de Simpson est :

$$\frac{(\beta - \gamma)^2}{2(\beta - \alpha)(\delta - \gamma)}$$

4. Généralisation du paradoxe de Simpson

Partons de la situation suivante :

$$X : \Omega \rightarrow \{x_1, \dots, x_m\}$$
$$Y : \Omega \rightarrow \{y_1, \dots, y_n\}$$
$$Z : \Omega \rightarrow \{z_1, \dots, z_l\}$$

Sur les espaces X et Y nous supposons qu'il existe des ordres totaux :

$$x_1 < \dots < x_m$$
$$y_1 < \dots < y_n$$

Ces ordres nous permettent d'introduire la notion de *Croissance Stochastique* que nous allons rappeler sommairement (cf : [5]).

Soit $\mathcal{X} = \{x_1, \dots, x_m\}$ muni d'un ordre total : $x_1 < \dots < x_m$

Sur l'ensemble des probabilités sur $\mathcal{X}$, nous pouvons définir *l'ordre du cumul* (ou *ordre stochastique*) par :

$$Q < P \Longleftrightarrow \sum_{i=k}^{i=m} Q_i \leqslant \sum_{i=k}^{i=m} P_i \text{ pour } k = 1, \dots, m$$

(Le sens de la relation $Q < P$ est : *P charge plus les grandes valeurs que Q*).

Si à une probabilité P, nous associons sa fonction de répartition D_P cumulée de *droite à gauche* :

$$D_P(x_k) = \sum_{i=k}^{i=m} P_i \quad \text{pour } k = 1, ..., m$$

nous avons :

$$Q < P \Longleftrightarrow D_Q \leqslant D_P$$

La figure (2) illustre cet ordre du cumul :

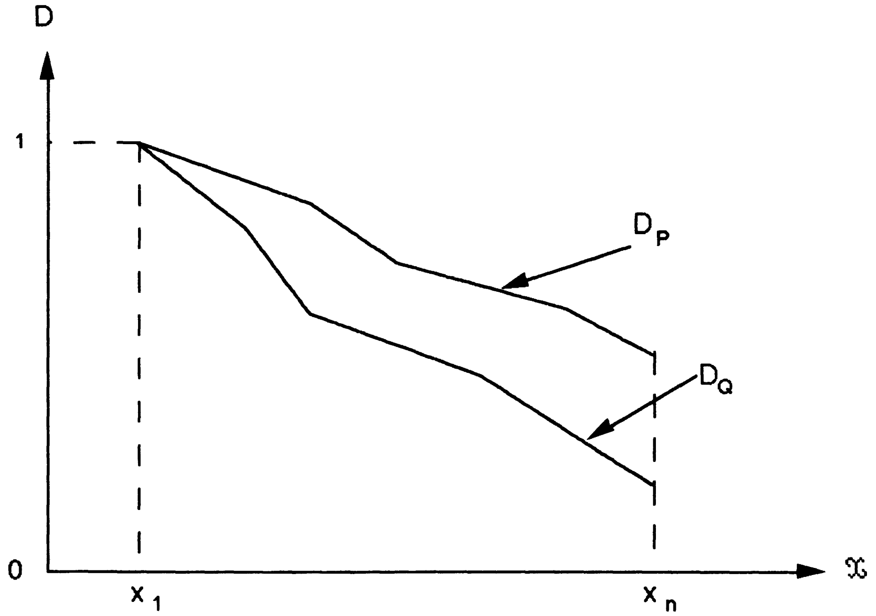

Figure 2
Représentation de l'ordre du cumul.

Cet ordre nous permet d'introduire la notion de croissance stochatisque entre 2 caractères ordonnés X et Y. Nous dirons que le caractère X est *croissant stochastiquement* en le caractère Y si la *transition conditionnelle* $P(X/Y)$ est une *transition croissante*, c'est-à-dire :

$y_\alpha < y_\beta \Longrightarrow P(X/y_\alpha) < P(X/y_\beta)$ (ordre du cumul sur les probabilités conditionnelles)

Cette notion formalise la phrase suivante : plus Y augmente, plus X augmente.

La figure (3) illustre cette croissante stochastique en utilisant les distributions conditionnelles cumulées de droite à gauche :

$$D(X/y_\alpha)(x_k) = \sum_{i=k}^{i=m} P(X = x_i/y_\alpha) \text{ pour } k = 1, \ldots, m$$

La généralisation du Paradoxe de Simpson consiste en le phénomène suivant :

La Croissance Stochastique de X par rapport à Y, Z étant fixé n'entraîne pas la croissance stochastique de X par rapport à Y.

Pour préciser cette notion, donnons un exemple de Paradoxe de Simpson généralisé. Dans la ville d'Anchourie, considérons maintenant la population Ω de

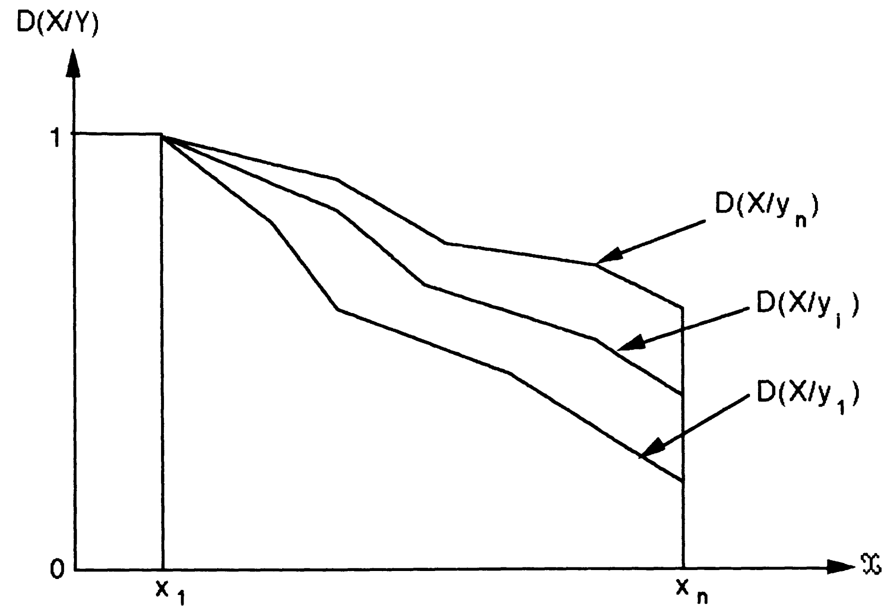

Figure 3
Croissance Stochastique de X en Y.

tous les lycéens des lycées Buffon et Condorcet :

$$L : \Omega \to \{b, c\}$$

Sur ces lycéens, nous étudions les deux caractères suivants :

$$X : \Omega \to \{x_1, x_2, x_3\} : \text{ Poursuite des Etudes Supérieures}$$
$$Y : \Omega \to \{y_1, y_2, y_3\} : \text{ Niveau d'Etudes du Père}$$

où

$x_1 < x_2 < x_3$

$x_1 =$ Non poursuite des Etudes Supérieures

$x_2 =$ Etudes Supérieures Courtes

$x_3 =$ Etudes Supérieures Longues

$y_1 < y_2 < y_3$

$y_1 =$ Niveau d'Etudes Bas

$y_2 =$ Niveau d'Etudes Moyen

$y_3 =$ Niveau d'Etudes Supérieur

Les tableaux (7) et (8) donnent les effectifs correspondant aux caractères X et Y pour les deux lycées :

Tableau 7
Effectifs $N(x, y, b)$
pour le lycée Buffon

Y / X	y_1	y_2	y_3
x_1	33	27	589
x_2	3	3	186
x_3	3	9	434

Tableau 8
Effectifs $N(x, y, c)$
pour le lycée Condorcet

Y / X	y_1	y_2	y_3
x_1	546	9	3
x_2	234	3	3
x_3	741	27	33

Dans chaque lycée, étudions les distributions conditionnelles $P(X/Y, L)$. Les tableaux (9) et (10) donnent les distributions conditionnelles cumulées de droite à gauche :

Tableau 9
Probabilités Conditionnelles Cumulées
$P(x/y, b)$ *pour le lycée Buffon*

$P(\ /y, b)$ / X	y_1	y_2	y_3
x_1	1	1	1
x_2	6/39	12/39	629/1209
x_3	3/39	9/39	434/1209

Tableau 10
Probabilités Conditionnelles Cumulées
$P(x/y, c)$ *pour le lycée Condorcet*

$P(\ /y, c)$ / X	y_1	y_2	y_3
x_1	1	1	1
x_2	975/1521	30/39	36/39
x_3	741/1521	27/39	33/39

Les figures (4) et (5) illustrent la croissance stochastique, dans chaque lycée, de X en Y : *dans chaque lycée, l'augmentation du niveau d'études du père entraîne l'augmentation de la poursuite des études.*

Etudions maintenant le cas de la ville entière. Le Tableau (11), somme des deux tableaux (7) et (8), donne les effectifs correspondants aux caractères X et Y pour l'ensemble de la ville d'Anchourie :

Tableau 11
Effectifs $N(x, y, .)$ *pour la ville d'Anchourie*

Y / X	y_1	y_2	y_3
x_1	579	36	592
x_2	237	6	189
x_3	744	36	467

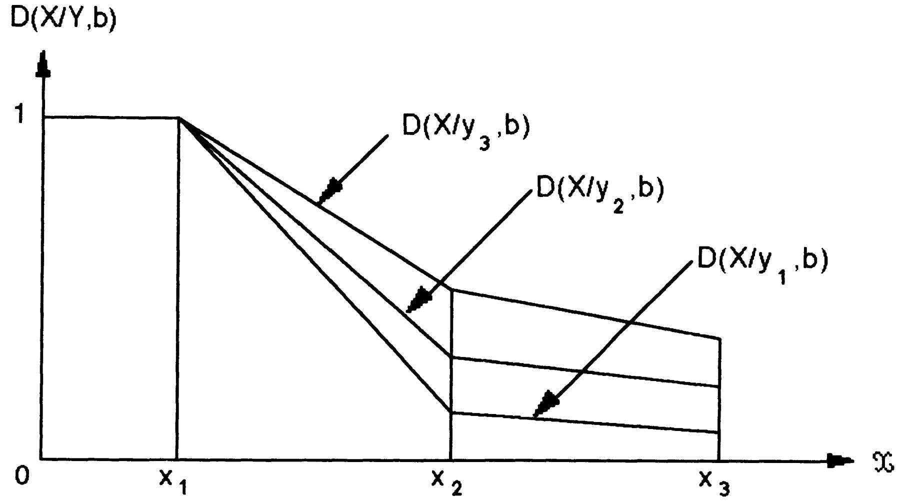

Figure 4
Croissance Stochastique de X en Y pour le lycée Buffon

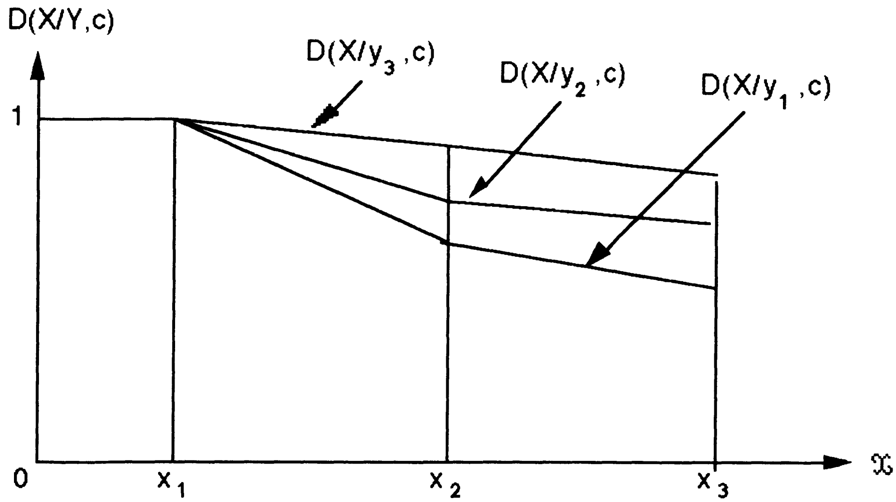

Figure 5
Croissance Stochastique de X en Y pour le lycée Condorcet

Le tableau 12 donne les distributions conditionnelles cumulées de droite à gauche :

Tableau 12
Probabilités Conditionnelles Cumulées $P(x/y)$
pour la ville d'Anchourie

X \ $P(\ /y)$	y_1	y_2	y_3
x_1	1	1	1
x_2	981/1560	42/78	656/1248
x_3	744/1560	36/78	467/1248

La figure (6), illustrant ce tableau, montre maintenant le phénomène inverse, une décroissance stochastique : *la diminution du niveau d'études du père entraîne l'augmentation de la poursuite des études.*

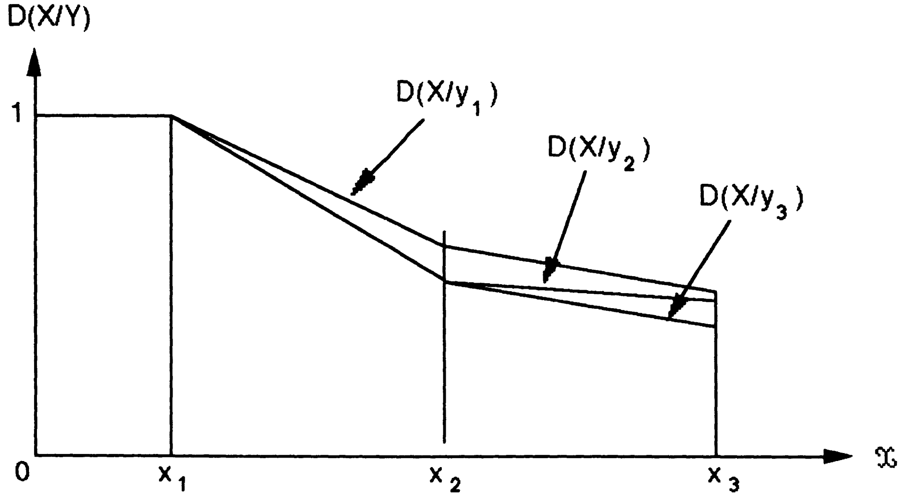

Figure 6
Décroissance Stochastique de X *en* Y *pour la ville d'Anchourie.*

Remarques :

1) La croissance stochastique que nous avons considérée est une notion multiplicative. On peut aussi définir une croissance stochastique additive qui ne produira pas l'effet Simpson. Nous pouvons faire, concernant cette croissance stochastique additive, les mêmes remarques qu'au paragraphe **1**. Si $X = \{x_1, \ldots, x_m\}$ est un espace muni d'un ordre total :

$$x_1 < \ldots < x_m$$

nous pouvons munir l'espace des mesures $\mathcal{M}(X)$ de deux structures d'ordres :

$$\nu \underset{d}{<} \mu \Longleftrightarrow \sum_{i=k}^{i=m} \nu(x_i) \leqslant \sum_{i=k}^{i=m} \mu(x_i) \text{ pour } k = 1, \ldots, m$$

$$\mu \underset{g}{<} \nu \Longleftrightarrow \sum_{i=1}^{i=k} \mu(x_i) \leqslant \sum_{i=1}^{i=k} \nu(x_i) \text{ pour } k = 1, \ldots, m$$

Si nous avons la condition de normalisation :

$$\nu(X) = \mu(X)$$

alors les deux structures d'ordres sont duales :

$$\nu \underset{d}{<} \mu \Longleftrightarrow \mu \underset{g}{<} \nu$$

2) Nous avons contruit cet exemple de Paradoxe de Simpson généralisé en utilisant les résultats du paragraphe III (en particulier la Figure (1)).

Conclusion :

Le Paradoxe de Simpson est une des illustrations de la phrase suivante, bien connue des statisticiens : «passer des effectifs aux fréquences n'est pas une opération sans danger».

Pour terminer, regrettons que le Paradoxe de Simpson (et sa généralisation) soit si peu étudié. Le classique Paradoxe de Condorcet a conduit à des études importantes (théorème d'Arrow par exemple) alors que le Paradoxe de Simpson est assez négligé. Il existe, dans la littérature statistique, des exemples réels de ce paradoxe.

Remerciements

L'auteur remercie le Professeur P. CAZES pour avoir relu son article et effectué des suggestions pour l'améliorer.

Bibliographie

[1] BLYTH C.R. (1972) "On Simpson's paradox and the sure thing principle" – J.A.S.A. – 67, 364-66.

[2] ENGEL A. (1975) "L'enseignement des probabilités et des statistiques" – CEDIC – Paris.

[3] HARDY C.H. and WRIGHT E.M.(1988) "An introduction to the theory of Numbers"-Clarenton Press-Oxford.

[4] OLIVIER J.P.(1985) "Barouf à Bombach" – Université P. Valéry – Montpellier.

[5] PETIT J.L.(1991) ”Polytopes convexes et structures d’ordre associés aux espaces de probabilités” – C.R.A.S. – E.N.S.P. – Yaoundé.

[6] SZEKELY G.J.(1986) ”Paradoxes in probability theory and mathematical statistics” – D. Reidel Pub. Comp. – Dordrecht.