

NOTE TO USERS

This reproduction is the best copy available.

UMI[®]

STEVE FECTEAU

PLANS DE SONDAGE ET ESTIMATION DE L'ERREUR
ÉCHANTILLONNALE DES INDICES DE PAUVRETÉ ET
D'INÉGALITÉ : UNE APPLICATION POUR LE BURKINA FASO

Mémoire

présenté

à la Faculté des études supérieures

de l'Université Laval

pour l'obtention

du grade de maître ès arts (M.A.)

Département d'économique

FACULTÉ DES SCIENCES SOCIALES

UNIVERSITÉ LAVAL

avril 2001

© Steve Fecteau, 2001



**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-60716-X

Canada

Résumé

Lorsqu'il s'agit de mesurer la pauvreté ou l'inégalité pour des fins de comparaisons, les économistes utilisent très souvent les indices de pauvreté et d'inégalité. Parmi les plus courants, on retrouve le niveau de vie per capita, les indices de FOSTER, GREER et THORBECKE et la dominance stochastique au niveau de la pauvreté, de même que l'indice d'ATKINSON pour l'inégalité. Afin d'estimer ces indices, on utilise de plus en plus les techniques d'enquête auprès des ménages. Dans la plupart des cas, les sondages ont des plans stratifiés, plusieurs degrés d'échantillonnage ainsi que des probabilités de sélection inégales pour au moins un des degrés d'échantillonnage. Le fait de négliger le plan de sondage, sous l'hypothèse d'un plan aléatoire simple, a des répercussions sur la qualité, soit le biais et la variance, des estimateurs. Il est donc primordial de prendre en compte le plan de sondage employé par les enquêteurs lorsque l'on désire mesurer la pauvreté et l'inégalité des individus au moyen d'une enquête auprès des ménages. Ces propos sont vérifiés empiriquement à l'aide des données d'une enquête statistique réalisée au Burkina Faso par l'Institut National de la Statistique et de la Démographie au cours de la période d'octobre 1994 à janvier 1995.

Avant-propos

Je tiens à remercier le Centre canadien d'étude et de coopération internationale (CECI) de même que le Centre de recherche en économie et finance appliquées pour le financement qui m'a permis de me consacrer entièrement au programme MIMAP-formation. Je remercie principalement mon directeur, monsieur Jean-Yves Duclos (professeur agrégé et co-responsable du projet), ainsi que mon employeur, monsieur Louis-Marie Asselin (directeur de la formation au CECI et également co-responsable du projet), pour la confiance et le support professionnel accordés. Je profite également de l'occasion pour exprimer ma gratitude envers messieurs Bernard Decaluwé, Luc Savard et John Cockburn pour m'avoir introduit au CRÉFA.

Comme il serait peu raisonnable de dresser une liste exhaustive des individus ayant contribué directement ou indirectement (de près ou de loin) à l'avancement de mes études de deuxième cycle, je me contenterai de nommer les principaux concernés dans l'espoir qu'ils auront apprécié autant que moi les moments passés ensemble. Je nomme et remercie sincèrement mes parents (Donald Fecteau et Monique Lachance), mon frère (Miguel Fecteau), mes lecteurs (madame Lynda Khalaf et monsieur Guy Lacroix), mon collègue Jimmy Royer (pour les cafés et les renseignements intéressants), ma collègue Anyck Dauphin (pour sa grande patience et ses encouragements), messieurs Denis Bolduc, Dimitri Sanga, Nicolas Beaulieu et Louis-Paul Rivest. Pareillement, tous ceux qui se sentent concernés par l'achèvement de mes études et qui me souhaitent succès et prospérité peuvent se considérer comme remerciés!

Dans un autre ordre d'idée, je voudrais rappeler que ce fut pour moi un plaisir d'oeuvrer au sein du CECI et du CRÉFA. Ce fut une expérience enrichissante. Cela m'a permis de rencontrer des gens fort sympathiques, dynamiques et d'en apprendre davantage à propos des pays en développement (PED). J'espère avoir un jour la chance de partager mes connaissances et mon savoir faire en analyse de la pauvreté avec des chercheur(e)s de l'étranger.

Table des matières

Résumé	i
Avant-propos	ii
1 Introduction	1
2 Problématique et théorie économique	4
2.1 Les indicateurs du niveau de vie	4
2.1.1 Les dépenses totales de consommation per capita	5
2.1.2 Les dépenses totales de consommation par équivalent adulte	6
2.2 Les seuils de pauvreté	7
2.2.1 Les seuils de pauvreté absolue	7
2.2.2 Les seuils de pauvreté relative	9
2.3 Les mesures de pauvreté	10
2.3.1 Le niveau moyen de vie réel per capita	10
2.3.2 La classe des indices FGT	10
2.4 La décomposition des indices de pauvreté	15
2.5 L'inégalité du niveau de vie	17
2.5.1 L'indice d'Atkinson	18

2.6	La dominance stochastique	20
2.6.1	La courbe d'incidence de la pauvreté	20
2.6.2	La courbe du déficit de la pauvreté	21
2.6.3	La courbe d'intensité de la pauvreté	24
3	Méthodologie et théorie statistique	25
3.1	Plan d'échantillonnage	26
3.1.1	Échantillonnage aléatoire simple avec probabilités de sélection égales	26
3.1.2	Échantillonnage avec probabilités de sélection inégales	27
3.1.3	Échantillonnage stratifié	31
3.1.4	Échantillonnage à plusieurs degrés	34
3.1.5	L'effet de plan	40
3.1.6	Les intervalles de confiance	40
3.2	Estimation des indices de pauvreté et d'inégalité	41
3.2.1	Les dépenses et la population	41
3.2.2	Le niveau moyen des dépenses réelles per capita	42
3.2.3	Les indices de pauvreté de la classe FGT	43
3.2.4	L'indice d'inégalité d'Atkinson	44
3.2.5	La dominance stochastique	46
4	Données de l'Enquête Prioritaire	48
5	Résultats et analyse empirique	54
5.1	Résultats sommaires	54
5.1.1	Le Burkina Faso	54
5.1.2	Les zones rurale et urbaine	55
5.1.3	Les groupes socio-économiques	57

5.1.4	Les strates d'analyse	59
5.2	Analyse statistique	60
5.2.1	La Pauvreté	61
5.2.2	Un domaine d'étude	61
5.2.3	Un profil de pauvreté	63
5.2.4	L'inégalité	64
5.2.5	La dominance stochastique	64
6	Conclusion	67
7	Bibliographie	70
A	Tableaux	73
B	Figures	83
C	Compléments théoriques	94
C.1	La linéarisation des estimateurs non linéaires	95
C.1.1	L'indice FGT	95
C.1.2	L'indice d'Atkinson	97
C.2	Estimation de la variance d'un estimateur de total	98
C.3	Comparaison de l'échantillonnage stratifié proportionnel à l'échantillonnage simple	99

Liste des tableaux

A.1	Le niveau de vie	74
A.2	Les indices FGT	76
A.3	L'indice d'ATKINSON	78
A.4	Tester l'écart entre deux courbes de dominance stochastique	80
A.5	Tester l'ordonnée d'une courbe de dominance stochastique	82

Liste des figures

1	Mesures de la pauvreté des individus	14
2	Intersection des courbes d'incidence de la pauvreté	23
B.1	Courbes de l'incidence de la pauvreté pour les huit strates d'analyse (20000 - 80000 f. CFA)	84
B.2	Courbes de l'incidence de la pauvreté pour les huit strates d'analyse (15000 - 30000 f. CFA)	85
B.3	Courbes de l'incidence de la pauvreté pour les sept groupes socio-économiques (20000 - 80000 f. CFA)	86
B.4	Courbes de l'incidence de la pauvreté pour les sept groupes socio-économiques (15000 - 30000 f. CFA)	87
B.5	Courbes de l'incidence de la pauvreté pour les deux zones géographiques (20000 - 80000 f. CFA)	88
B.6	Courbes de l'incidence de la pauvreté pour les deux zones géographiques (15000 - 30000 f. CFA)	89
B.7	Courbes du déficit de la pauvreté pour les groupes socio-économiques 6 et 7 (20000 - 80000 f. CFA)	90
B.8	Courbes du déficit de la pauvreté pour les groupes socio-économiques 6 et 7 (15000 - 30000 f. CFA)	91

B.9 Courbes de l'intensité de la pauvreté pour les groupes socio-économiques 6 et 7 (20000 - 80000 f. CFA)	92
B.10 Courbes de l'intensité de la pauvreté pour les groupes socio-économiques 6 et 7 (15000 - 30000 f. CFA)	93

Chapitre 1

Introduction

Il est important de pouvoir mesurer avec précision la pauvreté et l'inégalité. Cela permet d'établir avec confiance si elles sont plus élevées en un endroit plutôt qu'un autre, ou si elles se sont intensifiées au cours d'une période donnée. On peut alors mieux cibler certaines politiques de lutte contre la pauvreté et l'inégalité afin de vérifier l'impact social d'un ensemble de politiques économiques. Souvent, une comparaison ordinale de la pauvreté ou de l'inégalité sera suffisante lorsqu'il s'agit de comparer deux ou plusieurs situations ; c'est le cas, par exemple, lorsqu'il faut choisir entre deux actions gouvernementales. Les comparaisons cardinales exigeront que les différences d'intensité de la pauvreté ou d'inégalité soient quantifiées ; cela est utile lorsque l'on désire mesurer les effets précis d'une politique sur le niveau de pauvreté ou d'inégalité.

Afin de quantifier la pauvreté et l'inégalité, les chercheurs utilisent différents indices. Ces indices peuvent être représentés par des estimateurs statistiques calculés à partir d'un indicateur du niveau de vie et d'un seuil de pauvreté. Les différents indicateurs du niveau de vie employés dans la littérature sont un sujet controversé du fait qu'ils font souvent appel à certains jugements de valeur. Il en va de même pour les différentes méthodes visant à établir le seuil en deçà duquel un individu ou un ménage est considéré comme pauvre. Tout

comme les indices de pauvreté, les indices d'inégalité les plus courants sont obtenus à partir d'une mesure du bien-être de la population et peuvent être estimés par une fonction des données d'un échantillon.

De nos jours, les économistes-statisticiens ont souvent recours aux enquêtes auprès des ménages lorsqu'il s'agit d'estimer certains indices de pauvreté et d'inégalité. Les sondages permettent d'obtenir l'information directement des ménages et des individus qui les composent. L'information fournie par ces sondages se compare avantageusement à celle fournie par les agrégats économiques présentés dans les publications statistiques nationales.

L'objet de cette étude sera donc d'étudier des estimateurs de mesures de pauvreté et d'inégalité d'une population pour une ou plusieurs variables d'intérêt (par exemple les dépenses de consommation). De plus, nous présenterons les différents estimateurs de la variance, ou de l'erreur quadratique moyenne dans le cas de ratios de variables aléatoires, pour chacun de ces estimateurs. Ainsi, il sera possible d'estimer leur distribution empirique. Cela nous permettra ensuite d'effectuer des comparaisons de la pauvreté et de l'inégalité, de dresser différents profils de pauvreté, et d'établir statistiquement la précision de ces comparaisons et de ces profils.

La pertinence et l'originalité de ces recherches portent sur l'utilisation des techniques de sondage, ou plus précisément sur la prise en compte de la structure de l'enquête. Nous considérerons différents plans de sondage, allant de l'échantillon aléatoire simple (EAS) aux plans de sondage plus complexes, et nous analyserons les impacts de ces derniers sur la qualité des estimateurs (biais et variance). Nous verrons que les plans plus complexes avec stratification et probabilités de sélection inégales auront généralement tendance à améliorer la précision des estimateurs, tandis que les plans à plusieurs degrés produiront des estimateurs moins précis que les estimateurs de l'EAS, pour une taille d'échantillon fixe.

Par souci de simplicité, nous croyons qu'il est préférable de traiter le sujet selon deux approches distinctes. Dans un premier temps, nous aborderons l'aspect économique de la question, soit la problématique. Les thèmes étudiés seront les suivants : indicateurs du niveau de vie, seuils de pauvreté, indices de pauvreté, décomposition et profils de pauvreté, indice d'inégalité ainsi que la robustesse des comparaisons. Dans un deuxième temps, nous incorporerons l'aspect statistique et méthodologique. Ici, nous analyserons les plans de sondage généralement utilisés lors des enquêtes et présenterons les estimateurs utilisés dans le cadre du projet de recherche. Dans un troisième temps, nous exposerons et analyserons les résultats empiriques obtenus à partir des données de l'Enquête Prioritaire réalisée au Burkina Faso (d'octobre 1994 à janvier 1995) dans le cadre du Programme d'Ajustement Social (PAS). Ce programme fut mené conjointement par l'Institut National de la Statistique et de la Démographie (INSD) ainsi que par le Centre d'Études, de Documentation et de Recherche Économique et Sociale (CEDRES) avec l'appui de la Banque Mondiale et sous la supervision du Ministère de l'Économie, des Finances et du Plan.

Notre étude, réalisée dans le cadre du programme MIMAP («Micro-Impacts of Macroeconomic and Adjustment Policies»), dirigé par le CRÉFA (Centre de recherche en économie et finance appliquées) et le CECI-Québec (Centre canadien d'étude et de coopération internationale), consistera à utiliser les données déjà employées par l'INSD afin d'obtenir des estimations pour les indices de pauvreté et d'inégalité. De même, nous fournirons des estimations de la variance des estimateurs, tenant compte du plan de sondage qui fut utilisé lors de l'Enquête Prioritaire au Burkina Faso.

Chapitre 2

Problématique et théorie économique

2.1 Les indicateurs du niveau de vie

Le bien-être individuel peut être mesuré selon différentes méthodes. Il est important de noter que ces méthodes comportent souvent un aspect normatif, le bien-être étant multidimensionnel et subjectif. RAVALLION (1996), DEATON (1994) et SEN (1979) décrivent quelques indicateurs couramment utilisés. Parmi eux, la consommation courante et le revenu, l'unité d'observation étant le ménage ou l'individu, sont les plus utilisés. La consommation devrait tenir compte des produits alimentaires et non alimentaires, la valeur monétaire étant établie à partir des prix du marché autant que possible. De plus, elle devrait tenir compte de l'accessibilité aux biens et services publics même si leur valeur monétaire peut difficilement être établie. Il serait de même important de considérer toutes les sources de revenu lorsque l'approche revenu est retenue. Cela devrait inclure les revenus en nature, l'autoconsommation ainsi que la valeur des terres et du logement pour les propriétaires. Il est aussi très fréquent de diviser le montant des dépenses (ou le revenu) par un indice de prix propre au

lieu et à la période de sorte à obtenir une mesure standardisée, soit le niveau des dépenses réelles totales.

Le fait que le revenu soit plus volatile que la consommation (les revenus en milieu rural sont plutôt saisonniers) est une des raisons principales pour lesquelles il est souvent jugé préférable de retenir la consommation comme mesure du niveau de vie. Les ménages ont aussi tendance à *lisser*¹ leur consommation sous l'hypothèse du revenu permanent. Le choix de la consommation courante comme indicateur de bien-être peut toutefois porter à confusion. Même si la consommation est lissée, elle varie tout au long du cycle de vie. Par ailleurs, il est connu que les ménages moins nantis ont beaucoup plus de difficulté à emprunter, ce qui peut perturber le lissage de leur consommation.

Dans notre étude empirique, nous aurons à utiliser des mesures de bien-être. Ces mesures nous sont fournies par l'équipe de l'INSD déjà en place au Burkina Faso. Cette dernière a eu recours à deux méthodes fréquemment utilisées pour mesurer le niveau de vie des individus. Les deux méthodes sont les suivantes :

2.1.1 Les dépenses totales de consommation per capita

Cette mesure est en fait très simple. Il suffit d'estimer le montant des dépenses réelles totales d'un ménage et de diviser celui-ci par la taille du ménage. En utilisant cette méthode, on pose l'hypothèse que la consommation est également répartie entre les membres du ménage. Cette mesure ne tient pas compte des besoins variables de chacun des individus ainsi que des économies d'échelle² réalisables.

¹Le lissage de la consommation consiste à réduire la variation de la consommation le long du cycle de vie d'un individu. Ainsi, le niveau de consommation pour une période donnée suit d'assez près le revenu moyen de l'individu au cours de sa vie, plutôt que le revenu de la période.

²Le terme économie d'échelle signifie que les besoins totaux d'un ménage augmentent moins que proportionnellement au nombre d'individus.

2.1.2 Les dépenses totales de consommation par équivalent adulte

La consommation par équivalent adulte est égale au montant des dépenses réelles totales divisé par une échelle d'équivalence qui tient compte de la taille et de la composition du ménage. Le ménage composé d'un seul adulte est souvent utilisé comme ménage de référence. Cette méthode a pour avantage de prendre en compte les besoins des différents types d'individus (femme, homme ou enfant) ainsi que les économies d'échelle réalisables à l'intérieur d'un ménage.

Pour les fins de notre étude, nous aurons recours au niveau des dépenses réelles per capita (variable NVIE). Cet indicateur de niveau de vie est encore le plus employé dans la littérature étant donné que l'on a pas toujours accès à des mesures de composition telle l'*équivalent adulte*. De plus, les dépenses de consommation, exprimées en francs CFA, seront déflatées selon un indice de prix établi lors de l'Enquête Prioritaire d'octobre 1994. L'indice de la strate Ouagadougou–Bobo-Dioulasso servira de numéraire.

2.2 Les seuils de pauvreté

Maintenant que nous avons défini un indicateur nous permettant de mesurer le bien-être des individus, il reste à déterminer un (ou plusieurs) seuil de pauvreté. Le seuil de pauvreté représente un niveau de vie que l'individu doit avoir atteint pour ne pas être considéré comme pauvre. Il a pour fonction de répartir les ménages en deux classes (les pauvres et les non pauvres). Tout comme les indicateurs du niveau de vie, les seuils de pauvreté dépendent souvent de jugements normatifs, ce qui suscite souvent la controverse.

Nous présenterons diverses techniques visant à déterminer le seuil de pauvreté pour une région ou un sous-groupe de la population. Tout comme à la section précédente, la théorie sera présentée de façon simple et brève, l'objectif n'étant pas d'approfondir ces notions. Nous référons le lecteur à RAVALLION (1996), DEATON (1994) et ATKINSON (1987) pour une étude plus exhaustive.

2.2.1 Les seuils de pauvreté absolue

Un seuil de pauvreté absolue peut être perçu comme un seuil de «survie». Les seuils de survie sont toutefois très peu utilisés en pratique. Même dans les pays les plus pauvres, la subsistance n'est pas la seule préoccupation de l'être humain. Il serait plutôt préférable de considérer un seuil de pauvreté absolue comme étant une constante en terme du niveau de vie et unique dans le «domaine»³ où les comparaisons de la pauvreté sont effectuées. Deux individus ayant le même niveau de vie seront alors classés dans la même catégorie peu importe le lieu ou le moment. De plus, le seuil retenu doit respecter (tout comme l'incateur du niveau de vie) les objectifs de comparaison du domaine spécifié.

Les méthodes couramment utilisées débutent par la détermination de certains besoins de consommation de base, dont principalement la consommation de denrées alimentaires.

³Le domaine peut référer à une population, un groupe, un sous-groupe, une période, etc.

Ensuite, viennent les dépenses non alimentaires. RAVALLION et BIDANI (1994) proposent deux méthodes généralement appliquées en pratique (il en existe plusieurs autres) :

La méthode de l'énergie nutritive : «The Food Energy Intake (FEI) method»

Un concept souvent utilisé est celui de l'énergie nutritive. Ici, il faut estimer les dépenses de consommation (par individu ou équivalent adulte) qui permettent en moyenne d'obtenir un niveau d'énergie satisfaisant. Dans la pratique, on estime le coût moyen d'un ensemble de paniers consommés par des individus qui satisfont tout juste à leurs besoins nutritionnels. Il est également possible de régresser la consommation d'énergie nutritive par rapport aux dépenses de consommation totales et de déterminer le montant des dépenses correspondant au niveau d'énergie calorifique minimum. Cette méthode prend automatiquement en compte la consommation non alimentaire étant donné que l'on traite les dépenses totales. La relation entre dépenses et énergie ne sera sans doute pas la même pour chacune des régions. Il est à noter que cette méthode ne nécessite aucun indice de prix.

Le coût des besoins fondamentaux : «The Cost of Basic Needs (CBN) method»

Lorsque l'on a recours à cette méthode, il faut d'abord considérer un panier de référence qui satisfait les besoins élémentaires de l'individu. On peut estimer un seuil pour un individu moyen ou bien pour un «adulte moyen», selon l'indicateur choisi. Par la suite, on doit estimer le coût de ce panier. Un seuil peut être estimé pour chacun des sous-groupes qui doivent être comparés dans le profil de pauvreté. Le choix du panier pour chacun des sous-groupes doit normalement tenir compte de certains éléments : la zone (rurale ou urbaine), le niveau d'activité des individus, les habitudes de consommation (e.g. les traditions ancestrales), les prix relatifs pour les différentes régions, etc. Cette méthode a deux principaux inconvénients. D'une part, le panier de référence est construit de façon arbitraire et peut ne pas convenir à plusieurs ménages. D'autre part, il se peut que certains prix ne soient pas

disponibles. Cela se produit surtout dans le cas des services, des biens non alimentaires et des biens publics.⁴

2.2.2 Les seuils de pauvreté relative

On retrouve souvent le concept de pauvreté absolue dans l'étude des pays en développement, alors que les pays développés insisteront davantage sur le concept de pauvreté relative. Les seuils de pauvreté relative (SPR) sont souvent fixés par rapport à la moyenne ou la médiane de la distribution du revenu (e.g. 40 % de la moyenne nationale). Parfois, comme ce fut le cas pour l'Enquête Prioritaire (E.P.) au Burkina Faso, on fixe le seuil de pauvreté relatif à un certain quantile et le seuil d'extrême pauvreté relative à un quantile inférieur. Le seuil de pauvreté relative aura tendance à augmenter avec la croissance économique. Cette dernière définition de ligne de pauvreté ne sera pas utilisée dans ce mémoire.

Au cours de notre enquête portant sur le niveau de vie au Burkina Faso, nous utiliserons un seuil de pauvreté **absolue unique** par équivalent adulte obtenu selon la méthode du **coût des besoins fondamentaux**, avec prise en compte des besoins alimentaires spécifiques au Burkina Faso et des consommations non alimentaires. Un seuil absolu fut estimé à 41 099 f. CFA/an (estimé par l'INSD lors de l'Enquête Prioritaire d'octobre 1994). Nous dénoterons ce seuil par z . Un seuil absolu d'extrême pauvreté de 31 749 f. CFA/an fut aussi estimé.⁵ Nous poserons l'hypothèse forte que les besoins de chacun des membres du ménage sont identiques afin de pouvoir discriminer les ménages pauvres des ménages non pauvres à partir du niveau des dépenses réelles per capita.

⁴Il est à noter qu'il existe certaines variantes pour cette méthode quant à la détermination des dépenses non alimentaires requises.

⁵Le dollar canadien (\$) vaut présentement environ 400 francs CFA.

2.3 Les mesures de pauvreté

2.3.1 Le niveau moyen de vie réel per capita

Il s'agit d'une mesure standard du bien-être agrégé lorsqu'il est question de comparer deux populations ou deux sous-groupes d'une population. Le niveau moyen de vie réel per capita se définit simplement comme le niveau moyen des dépenses ou de consommation par individu. Il s'exprime ainsi :

$$\bar{Y} = \frac{Y_{total}}{N} = \frac{\sum_{j=1}^M T_j \cdot Y_j}{\sum_{j=1}^M T_j}, \quad (2.1)$$

où Y_j représente le niveau de vie réel per capita (variable d'intérêt) du ménage j , M le nombre de ménages (unités statistiques ou unités d'observation) de la population, T_j la taille du $j^{ième}$ ménage et N le nombre total d'individus dans la population.

2.3.2 La classe des indices FGT

Maintenant que nous avons un indicateur du niveau de vie et un seuil de pauvreté, il est possible de définir une mesure de la pauvreté individuelle. En agrégeant les mesures individuelles, nous obtenons un indice de pauvreté pour le domaine étudié. Le seuil de pauvreté sera considéré comme fixe et seuls les indices de pauvreté respectant le principe de l'additivité seront considérés. Ce dernier concept sera défini plus loin.

Tel que cité par SEN (1976), il serait souhaitable que les indices utilisés lors d'une analyse de pauvreté respectent les deux axiomes fondamentaux qui suivent :

Axiome de monotonie : Toutes autres choses étant égales, une diminution du bien-être d'une personne sous le seuil de pauvreté doit augmenter la mesure de pauvreté.

Axiome des transferts : Toutes autres choses étant égales, un pur transfert de bien-être d'une personne sous le seuil de pauvreté vers une autre personne ayant un niveau de bien-être plus élevé (pauvre ou pas) doit augmenter la mesure de pauvreté (aussi connu sous le «principe de Pigou-Dalton»).

L'indice numérique de pauvreté

Soit q le nombre d'individus ayant un niveau de bien-être inférieur au seuil de pauvreté et N le nombre total d'individus de la population.

$$H = \frac{q}{N} = \frac{\sum_{j=1}^M T_j \mathbb{k}(Y_j \leq z)}{\sum_{j=1}^M T_j}, \quad (2.2)$$

où $\mathbb{k}(Y_j \leq z)$ est une fonction indicatrice prenant la valeur 1 lorsque l'inégalité est respectée et 0 sinon. Les individus non pauvres ont donc un poids statistique nul dans cette mesure et il en sera de même pour les autres mesures.

Cet indice est le plus simple et le plus employé dans la littérature. Il indique le pourcentage de la population sous le seuil de pauvreté retenu. Il est grandement apprécié du fait qu'il est très simple à interpréter. Cet indice présente toutefois un grave inconvénient. Il ne tient aucunement compte du degré de pauvreté des individus. Cet indice ne respecte aucun des deux axiomes. Supposons, par exemple, qu'un individu pauvre s'appauvrisse. Il est évident que la pauvreté s'est accrue, *ceteris paribus*, mais l'indice numérique est invariant.

L'écart (relatif) de pauvreté

$$PG = \frac{\sum_{j=1}^M T_j \left(\frac{z-Y_j}{z} \right) \mathbf{k}(Y_j \leq z)}{\sum_{j=1}^M T_j}, \quad (2.3)$$

où $\left(\frac{z-Y_j}{z} \right)$ représente la mesure de pauvreté individuelle.

Cet indicateur est aussi très facile à interpréter. Il exprime l'écart (normalisé par z) moyen qui sépare les pauvres du seuil de pauvreté. Il donne une idée de l'intensité de la pauvreté des individus. Il respecte l'axiome de monotonie, mais pas l'axiome des transferts. Cet indice est sensible aux transferts effectués entre les pauvres et les non pauvres mais n'est malheureusement pas sensible aux transferts entre les individus pauvres et demeurant pauvres par la suite (la somme des écarts demeure constante dû à la linéarité de la mesure individuelle).

Afin de satisfaire les deux axiomes de SEN, l'indice suivant fut proposé par FOSTER, GREER et THORBECKE (1984) :

La sévérité de la pauvreté

$$P_2 = \frac{\sum_{j=1}^M T_j \left(\frac{z-Y_j}{z} \right)^2 \mathbf{k}(Y_j \leq z)}{\sum_{j=1}^M T_j}. \quad (2.4)$$

Cet indice respecte les deux axiomes. Il est affecté par le niveau de pauvreté des individus. De plus, il est sensible à la direction dans laquelle s'effectue un transfert. La valeur

prise par P_2 s'accroîtra si un transfert a pour effet d'appauvrir un individu pauvre pour enrichir un individu plus riche, même si ce dernier demeure pauvre après le transfert.

Les indices de pauvreté présentés ci-haut ont une mesure commune. On peut les regrouper dans une même classe, celle des indices **FGT** ou **P-Alpha**. Leur forme générale se présente ainsi :

$$P_\alpha = \frac{\sum_{j=1}^M T_j \left(\frac{z-Y_j}{z}\right)^\alpha \mathbb{k}(Y_j \leq z)}{\sum_{j=1}^M T_j}, \quad (2.5)$$

où α est un paramètre d'aversion à la pauvreté ($\alpha \geq 0$ et $\alpha \in N$).

Lorsque $\alpha = 0$, on retrouve H (P_0). Si $\alpha = 1$, il s'agit de PG (P_1). Pour $\alpha > 0$, la mesure de pauvreté individuelle est une fonction strictement décroissante par rapport au niveau de vie et strictement croissante par rapport au seuil de pauvreté. Pour $\alpha > 1$, l'indice de pauvreté est également sensible aux transferts entre les pauvres (les deux axiomes sont satisfaits par la stricte convexité de la mesure individuelle $\left(\frac{z-Y_j}{z}\right)^\alpha$). Pour $\alpha \rightarrow \infty$, l'indice ne représente plus que la pauvreté de l'individu le plus pauvre (indice de RAWLS).

La figure 1 présente les mesures de la pauvreté individuelle selon les différents indices utilisés. Pour $\alpha > 0$ et assez grand, la discontinuité de $\left(\frac{z-Y_j}{z}\right)^\alpha$ en z tend à disparaître. Cette propriété est souhaitée en ce sens que le poids d'un individu dans la mesure de l'indice tend vers zéro lorsque son bien-être s'approche du seuil de pauvreté. Le fait de voir le bien-être d'un individu passer de $z - \varepsilon$ à $z + \varepsilon$, pour $\varepsilon > 0$ et petit, n'a sensiblement pas d'effet sur l'indice de pauvreté si α est suffisamment grand.

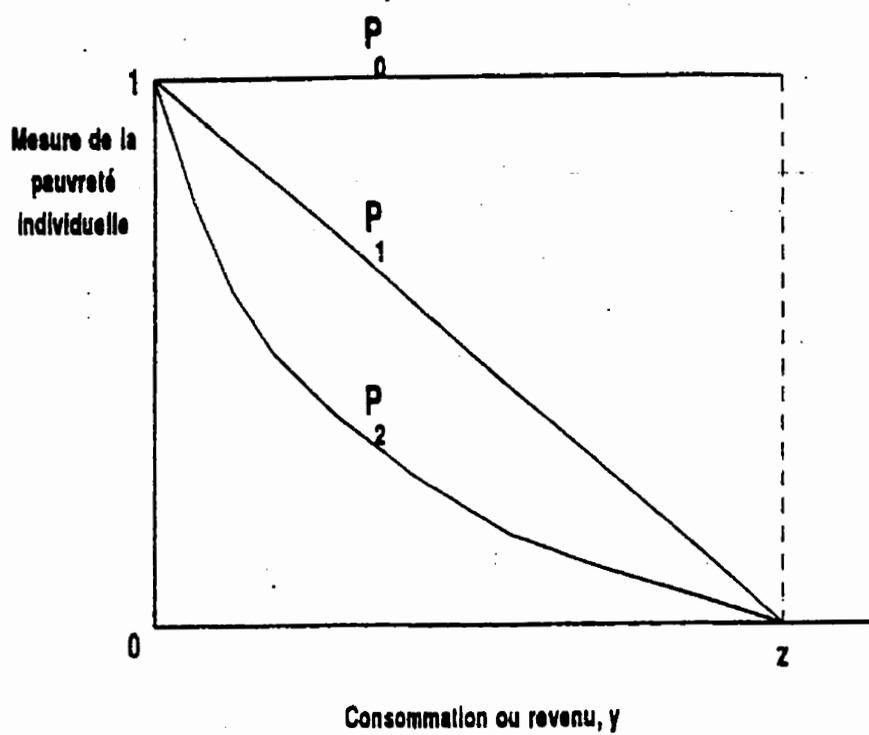


Figure 1
Mesures de la pauvreté des individus

Source : RAVALLION, M. (1996)

2.4 La décomposition des indices de pauvreté

Les indices additifs peuvent être décomposés en sous-indices, chacun mesurant la pauvreté d'un sous-groupe de la population. Soit G sous-groupes mutuellement exclusifs ; le **profil de pauvreté** est alors la liste des sous-indices de pauvreté, P_α^g , pour chacun des sous-groupes, $g = 1, \dots, G$. L'indice de pauvreté totale est tout simplement la somme pondérée des «sous-indices».

$$\begin{aligned}
 P_\alpha &= \frac{\sum_{g=1}^G \left(\sum_{j=1}^{M_g} T_{gj} \left(\frac{z - Y_{gj}}{z} \right)^\alpha \mathbf{k}(Y_{gj} \leq z) \right)}{\sum_{g=1}^G \sum_{j=1}^{M_g} T_{gj}} \\
 &= \frac{1}{N} \sum_{g=1}^G N_g P_\alpha^g = \sum_{g=1}^G \underbrace{\left(\frac{N_g}{N} \right)}_{\text{Contribution du groupe } g (C_\alpha^g)} P_\alpha^g,
 \end{aligned} \tag{2.6}$$

Contribution du groupe g (C_α^g)

où T_{gj} et Y_{gj} correspondent à la taille et au niveau de vie réel per capita du $j^{\text{ième}}$ ménage du $g^{\text{ième}}$ sous-groupe. De même, N_g/N correspond au poids (proportion dans la population) du $g^{\text{ième}}$ sous-groupe dans l'indice «global» de pauvreté.

L'avantage de la propriété de l'additivité est qu'elle garantit le respect de l'axiome de «cohérence» des sous-groupes. Lorsque la pauvreté s'accroît (diminue) dans un sous-groupe quelconque, *ceteris paribus*, la pauvreté globale ne peut diminuer (augmenter). Cette propriété est fort désirable lorsque l'on a pour objectif de réduire la pauvreté par mesures ciblées. En effet, si l'indicateur n'a pas cette propriété, il est possible de se trouver dans une situation où une politique locale réduit la mesure de pauvreté de la région ciblée tout en augmentant la mesure de pauvreté globale. Dans un pareil cas, il serait difficile de contrôler et de mesurer l'efficacité de cette politique. La propriété de cohérence des sous-groupes devrait être considérée comme essentielle si on veut que notre programme soit cohérent [FOSTER et SHORROCKS (1988)].

À la sous-section intitulée «**Analyse statistique**», nous présenterons un exemple de la décomposition d'un indice de pauvreté estimé à l'aide des données de l'enquête.

2.5 L'inégalité du niveau de vie

Le concept d'«inégalité» s'interprète de multiples façons. Ce concept fait généralement référence à un manque d'*égalité* concernant une dotation ou un traitement quelconque, comme la santé des individus, leur richesse, le prestige ou l'importance qui leur est accordée. Nous parlerons ici de l'inégalité comme étant un manque d'égalité au niveau du bien-être des individus. La mesure de bien-être utilisée sera, comme précédemment, le niveau des dépenses réelles per capita.

L'ampleur de l'inégalité peut se présenter selon deux approches. La première approche, plus visuelle et intuitive, consiste à donner une image de la distribution du niveau de vie de la population. Très souvent, on emploie la courbe de LORENZ, les parts du revenu, les quantiles, la distribution empirique, etc. La seconde consiste à fournir des mesures numériques de l'inégalité. Ces mesures, unidimensionnelles ou multidimensionnelles, permettent de comparer de façon ordinale le niveau d'inégalité entre deux situations.

Nous nous limiterons ici à une seule mesure d'inégalité. Il s'agit de l'indice d'ATKINSON [ATKINSON (1970), COWELL (1995) et DEATON (1994)]. Cet indice repose sur le concept de la fonction de bien-être social introduit initialement par DALTON en 1920 et reprise par ATKINSON (1970). L'originalité et la pertinence de ce mémoire découlera de l'estimation de la variance de cet estimateur non linéaire.

2.5.1 L'indice d'Atkinson

Considérant le niveau de vie moyen (\bar{Y}) défini à l'équation (2.1), l'indice d'inégalité d'ATKINSON ($0 \leq A_\varepsilon \leq 1$) s'exprime ainsi :

$$A_\varepsilon = 1 - \left[\frac{1}{\sum_{j=1}^M T_j} \sum_{j=1}^M T_j \left(\frac{Y_j}{\bar{Y}} \right)^{1-\varepsilon} \right]^{\frac{1}{1-\varepsilon}} \quad (2.7)$$

$$= 1 - \frac{1}{\bar{Y}} \cdot \left[\frac{\sum_{j=1}^M T_j (Y_j)^{1-\varepsilon}}{\sum_{j=1}^M T_j} \right]^{\frac{1}{1-\varepsilon}}, \quad \varepsilon \geq 0, \varepsilon \neq 1 \quad (2.8)$$

$$A_\varepsilon = 1 - \prod_{j=1}^M \left(\frac{Y_j}{\bar{Y}} \right)^{\frac{T_j}{\sum T_j}}, \quad \varepsilon = 1 \quad (2.9)$$

où ε est un coefficient d'aversion pour l'inégalité ($\varepsilon \geq 0$ et $\varepsilon \in \mathbb{N}$).

Dans le cas où $\varepsilon = 0$, l'indice d'inégalité prend la valeur nulle. Pour $\varepsilon = 1$, il s'agit de l'indice de BERNOULLI. Finalement, pour le cas limite où $\varepsilon \rightarrow \infty$, on retrouve l'indice de RAWLS.

Le fait d'augmenter le bien-être de tous dans une même proportion (ou d'augmenter le bien-être total sans modifier les parts de chacun) n'affecte pas l'indice d'inégalité d'ATKINSON. Pareillement, le fait de regrouper plusieurs populations identiques laisse l'indice invariant. Il est donc indépendant de l'échelle du bien-être total et de l'échelle de la population.

Pour les fins de l'étude sur l'inégalité au Burkina Faso, nous nous limiterons à l'estimation de l'indice d'ATKINSON ainsi que son erreur échantillonnale (ou l'erreur-type). Si nous avons choisi cet indice en particulier, c'est tout d'abord pour sa popularité en sciences économiques. Le coefficient de GINI ainsi que l'indice de THEIL sont aussi très utilisés dans la littérature économique. D'autres indices d'inégalité sont beaucoup plus simples, mais ne

reposit que sur le concept statistique de la variance (e.g. la variance, le coefficient de variation, la mesure exponentielle, etc.). Ils sont plus difficilement interprétables du point de vue normatif.

2.6 La dominance stochastique

Dans les comparaisons de pauvreté, il peut être préférable d'employer plus d'un seuil de pauvreté ou de considérer un intervalle de seuils lorsque nous ne pouvons les estimer avec certitude ou lorsque leur choix suscite la controverse. Les seuils sont souvent établis à partir de critères subjectifs, voir même de façon arbitraire. Le concept de dominance stochastique permet de tester la robustesse des comparaisons ordinales de la pauvreté lorsqu'il y a incertitude sur la valeur du seuil de pauvreté. Il permet aussi de faire une vérification simultanée sur les classes d'indices de pauvreté qui sont cohérentes avec un ordre particulier de dominance (voir FOSTER et SHORROCKS (1988)).

2.6.1 La courbe d'incidence de la pauvreté

Si on trace le graphe de $H = q/N$ en fonction du seuil de pauvreté z , on obtient la courbe d'incidence de la pauvreté. Cette courbe exprime le pourcentage d'individus pauvres étant donné un seuil de pauvreté quelconque. Il s'agit en fait de la fonction de distribution cumulée du niveau de vie $[F(z)]$. Pour obtenir cette courbe il suffit de tracer le graphe de P_0 en fonction de z , où $P_0(z)$ est défini comme suit :

$$P_0(z) = \int_0^z \left(\frac{z-y}{z}\right)^0 f(y) dy = F(z), \quad (2.10)$$

$$z \in [0, z_{\max}].$$

La courbe d'incidence de la pauvreté possède des propriétés intéressantes. En effet, si la courbe $F(z)$ de la situation A est supérieure à celle de la situation B pour tout seuil de pauvreté considéré, alors on peut affirmer que la pauvreté a clairement diminué de la situation A à la situation B pour tous les indices de pauvreté du premier ordre (ceux qui respectent le premier axiome de SEN). On est alors en présence de la Condition de dominance

du premier ordre. Lorsque les courbes se croisent, le classement entre les deux situations devient ambigu. On peut parfois réduire l'intervalle de variation du seuil de pauvreté ou vérifier la Dominance du second ordre.

2.6.2 La courbe du déficit de la pauvreté

La courbe du déficit de la pauvreté ($D(z)$) est obtenue en reportant, sur un deuxième graphe, la surface sous la courbe de l'incidence de la pauvreté pour chacun des seuils considérés. Mathématiquement :

$$\begin{aligned} P_1(z) &= \int_0^z \left(\frac{z-y}{z}\right)^1 f(y) dy = \frac{1}{z} \int_0^z (z-y) f(y) dy \\ &= \frac{1}{z} \int_0^z F(y) dy. \end{aligned} \quad (2.11)$$

$$D(z) = z \cdot P_1(z) = \int_0^z F(y) dy. \quad (2.12)$$

Cette courbe s'obtient directement à partir de P_1 (reporter $z \cdot P_1(z)$ en fonction de z).

La Condition de dominance du second ordre sera respectée lorsque la courbe du déficit de la pauvreté pour une situation B , sera inférieure à celle d'une autre situation (A) sur l'intervalle $z \in [0, z_{\max}]$. On peut alors affirmer que la situation B sera préférée à la situation A en terme de pauvreté pour tous ces seuils et pour tous les indices de pauvreté du deuxième ordre (ceux qui respectent les deux axiomes de SEN). La figure 2 présente les courbes de l'incidence de la pauvreté ainsi que les courbes du déficit de la pauvreté pour ces deux situations. Si le seuil de pauvreté maximum est inférieur à z^* , alors la pauvreté est plus intense dans la situation A (valable pour les trois indices FGT). C'est ce que démontre le diagramme a); nous sommes donc en présence de la Condition de dominance du premier ordre. Si le seuil de pauvreté admissible peut atteindre $z^{\max} > z^*$, alors la pauvreté sera plus intense dans la situation A pour tous les indices du deuxième ordre si la courbe du déficit

de pauvreté de A est supérieur à celle de B en tout point $z \in [0, z_{\max}]$. Ceci est valable pour les indices P_1 et P_2 , mais pas pour P_0 car nous sommes en présence de la Condition de dominance du second ordre.

Dans le cas où les courbes $D(z)$ se croisent, le classement devient ambigu. Il peut encore être possible de réduire l'intervalle pour z ou de procéder à un test du troisième ordre.

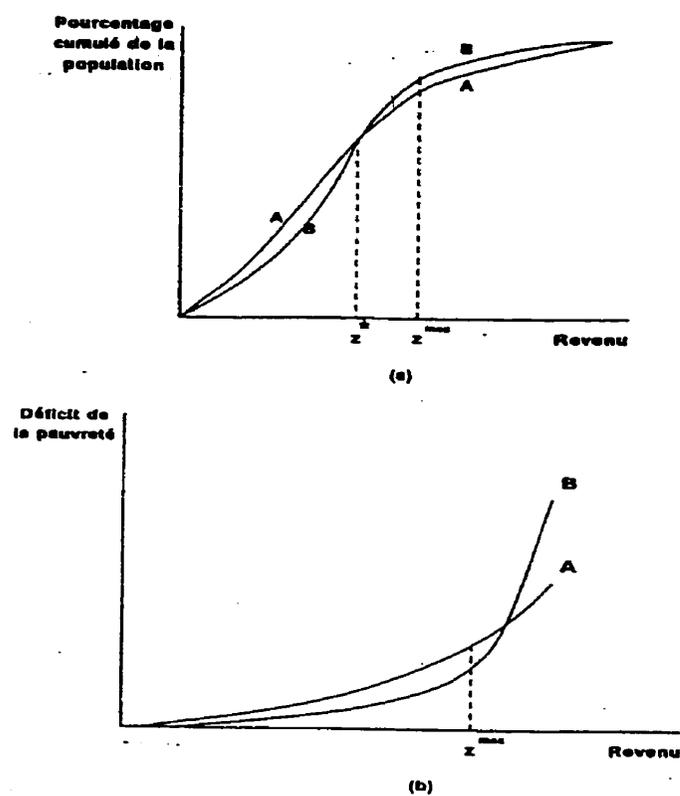


Figure 2

Intersection des courbes d'incidence de la pauvreté

Source : RAVALLION, M. (1996).

2.6.3 La courbe d'intensité de la pauvreté

Cette dernière courbe ($S(z)$) s'obtient de la même façon que la précédente, soit à partir de la surface sous la courbe du déficit de la pauvreté. Mathématiquement :

$$\begin{aligned} P_2(z) &= \int_0^z \left(\frac{z-y}{z}\right)^2 f(y) dy \\ &= \frac{1}{z^2} \int_0^z (z-y)^2 f(y) dy \end{aligned} \quad (2.13)$$

$$\begin{aligned} S(z) &= \frac{z^2}{2} \cdot P_2(z) \\ &= \int_0^z D(y) dy. \end{aligned} \quad (2.14)$$

La courbe de l'intensité de la pauvreté se trace assez facilement à l'aide de l'indice P_2 ($\frac{z^2}{2} \cdot P_2(z)$ en fonction de z).

Pour pouvoir comparer deux situations sans ambiguïté, il est nécessaire que l'une des deux courbes d'intensité de la pauvreté soit au dessus de l'autre pour tout $z \in [0, z_{\max}]$. On est alors en présence de la Condition de dominance du troisième ordre. Advenant le cas où il y aurait un croisement des courbes d'intensité de la pauvreté, il est toujours possible de tester une condition d'ordre plus élevée, bien que l'interprétation des mesures plus restrictives (e.g. P_3 ou P_4) devienne plus difficile.

Pour en savoir davantage à propos de la dominance stochastique, on peut se référer à ATKINSON (1987), DEATON (1994), DUCLOS et DAVIDSON (2000), FOSTER et SHORROCKS (1988), et RAVALLION (1996).

Chapitre 3

Méthodologie et théorie statistique

À la section précédente, nous avons défini une classe d'indices de pauvreté et d'inégalité. Afin d'estimer ces indices, nous aurons recours aux enquêtes auprès des ménages. DEATON (1994), HOWES et LANJOUW (1997,1998) ont contribué à l'application des techniques de sondage en sciences économiques, et plus particulièrement dans le domaine de l'estimation de la variabilité des indices de pauvreté et d'inégalité en présence de plans de sondage complexes. Bien que la théorie échantillonnale existe depuis des dizaines, voir même une centaine d'années, plusieurs chercheurs ne prennent pas en compte la structure réelle de l'enquête et se limitent à la formulation d'estimateurs provenant d'un plan aléatoire simple. L'objectif premier sera de montrer intuitivement et empiriquement l'impact du plan de sondage sur l'estimation de la variance des estimateurs statistiques. Nous présenterons initialement l'échantillonnage aléatoire simple (EAS) avec probabilités de sélection égales. Ce plan est le plus simple et on suppose qu'une base de sondage (e.g. liste électorale, bottin téléphonique, etc.) pour l'ensemble de la population est disponible (la taille de la population M est donc connue). Par la suite, nous ajouterons graduellement de la complexité au plan de sondage (probabilités de sélection inégales, la stratification, l'échantillonnage à deux degrés et les plans complexes) de sorte à ce que celui-ci soit un peu plus réaliste.¹ Le lecteur déjà initié

¹Les formules statistiques utilisées proviennent de ASSELIN (1984), de DEATON (1994), de COCHRAN (1977) et du U.S Bureau of Census (1995). Nous invitons le lecteur à consulter le guide de SATIN A. et SHASTRY

aux techniques échantillonnales peut passer directement à la section 3.2.

3.1 Plan d'échantillonnage

3.1.1 Échantillonnage aléatoire simple avec probabilités de sélection égales

Soit une population de M unités échantillonnales.² Nous tirons un échantillon de m unités sans remise (SR) à partir duquel nous mesurons x_j .³ Il est à noter que l'on ne fait aucune hypothèse sur la distribution des x_j ; ils ne sont pas nécessairement indépendants et identiquement distribués (*iid*). Toutefois, nous supposerons que les unités sont tirées avec remise (AR). La probabilité qu'une unité j soit échantillonnée est de $\pi_j \simeq \frac{m}{M} \forall j$ dans les cas où m est petit par rapport à M . On peut estimer $X_{total} = X = \sum_{j=1}^M X_j$ de façon non biaisée par \hat{X} :

$$\hat{X} = \sum_{j=1}^m \frac{x_j}{\pi_j} = \frac{M}{m} \sum_{j=1}^m x_j \quad (3.15)$$

$$V(\hat{X}) = M^2 \frac{\sigma^2}{m} \quad (3.16)$$

$$v(\hat{X}) = M^2 \frac{s^2}{m}, \quad (3.17)$$

où $V(\hat{X})$ représente la variance théorique et $v(\hat{X})$ est l'estimation non biaisée de la variance. Le fait de supposer que le tirage fut réalisé avec remise, bien que ce n'est pas le cas en réalité, simplifie quelque peu les formules pour la variance. Dans le cas d'un échantillonnage sans remise (SR), on doit multiplier la variance et son estimateur par un facteur de correction

W. (1993) : L'échantillonnage : un guide non mathématique, deuxième édition, Statistique Canada, Cat. n. 12-602F, 100p.

²Pour plus de simplicité, nous utiliserons le terme unité. Ceci évitera les risques de confusion entre ménages et individus. La distinction entre ces deux termes sera faite subséquemment.

³La variable X_j , $j = 1, \dots, M$ représente une variable d'intérêt quelconque pour la population alors que x_j , $j = 1, \dots, m$ est l'analogue dans le cas de l'échantillon.

pour population finie ($\frac{M-m}{M}$). Lorsque la fraction de sondage ($\frac{m}{M}$) est négligeable, il y a peu de différence entre un sondage avec ou sans remise. On voit assez facilement que le fait de considérer un sondage avec remise aura tendance à surestimer la variance de l'estimateur ($\frac{m}{M} > 0 \Rightarrow \frac{M-m}{M} < 1$). Lorsqu'il s'agit d'estimer une proportion, il suffit de poser $x_j = 1$ si l'unité appartient au domaine d'étude et $x_j = 0$ sinon. Un estimateur de total ou de ratio peut toujours s'interpréter comme une somme pondérée des variables de l'échantillon. Dans le cas de l'EAS, le poids de chacune des unités échantillonnées est $w = \pi^{-1} \simeq \frac{M}{m}$ pour tout j . La notion de poids échantillonnaux sera expliquée davantage à la prochaine sous-section.

Dans le cas d'une moyenne pour la population (\bar{X}), l'estimateur non biaisé s'obtient tout simplement en divisant l'estimateur du total \widehat{X} par la taille de la population M (constant) :

$$\widehat{X} = \frac{1}{M} \sum_{j=1}^m \frac{x_j}{\pi_j} = \frac{1}{m} \sum_{j=1}^m x_j = \bar{x} \quad (3.18)$$

$$V(\widehat{X}) = \frac{\sigma^2}{m} \quad (3.19)$$

$$v(\widehat{X}) = \frac{s^2}{m}. \quad (3.20)$$

Il est à noter que :

$$\sigma^2 = \frac{\sum_{j=1}^M (X_j - \bar{X})^2}{M}, \quad (3.21)$$

$$s^2 = \frac{\sum_{j=1}^m (x_j - \bar{x})^2}{m-1}. \quad (3.22)$$

3.1.2 Échantillonnage avec probabilités de sélection inégales

Pour l'EAS, la probabilité d'appartenir à l'échantillon ($\frac{m}{M}$) était la même pour chacune des unités de la population. Maintenant, nous poserons $\pi_j \simeq m p_j$ la probabilité pour l'unité

j d'appartenir à l'échantillon, où p_j est la probabilité que j soit sélectionné à un tirage particulier : ⁴

$$\sum_{j=1}^M \pi_j \simeq m \quad (3.23)$$

$$\sum_{j=1}^M p_j \simeq 1. \quad (3.24)$$

Dans le cas de l'EAS, $p_j = \frac{1}{M}$. L'estimateur utilisé sera l'estimateur de HORVITZ-THOMPSON (*HT*) :

$$\widehat{X}_{HT} = \sum_{j=1}^m \frac{x_j}{\pi_j}. \quad (3.25)$$

On a :

$$E(\widehat{X}_{HT}) = X. \quad (3.26)$$

Ici nous ne considérerons que la formulation avec remise (AR) puisqu'il sera plus facile d'obtenir une estimation de la variance. La valeur de la variance avec remise sera conservatrice.⁵

De même, il est possible de définir le poids échantillonnal, w_j , d'une unité échantillonnale (x_j) comme étant l'inverse de sa probabilité de sélection : $w_j \simeq (mp_j)^{-1}$. Lorsque l'on ne connaît pas le nombre d'unités dans la population, il est possible de l'estimer par $\widehat{M} = \sum_{j=1}^m w_j$. Dans le cas de l'EAS (le nombre d'unités est connu, M), $w_j = \frac{1}{\pi_j} = \frac{M}{m}$

⁴Dans notre étude, les probabilités de sélection furent calculées par l'INSD lors de l'Enquête Prioritaire de 1994. Nous référons le lecteur au rapport méthodologique [INSD (1996)].

⁵Lorsque nous utilisons le terme «conservatrice», nous voulons dire que la variance calculée surestime la vraie variance. Il est conservateur de ne pas surestimer la précision d'un estimateur.

$\Rightarrow \widehat{M} = \sum_{j=1}^m \frac{M}{m} = m \times \frac{M}{m} = M$. Le poids d'une unité j peut être interprété comme étant le nombre d'unités que représente cette unité échantillonnée.

Pour les différents plans de sondage décrits dans ce mémoire, on a toujours $E(\widehat{M}) = M$ pour \widehat{M} défini précédemment. De même, X peut être estimé par $\widehat{X} = \sum_{j=1}^m w_j x_j$ et $E(\widehat{X}) = X$.

La statistique recherchée est la suivante :

$$V(\widehat{X}) = \frac{1}{m} \sum_{j=1}^m \pi_j \left(\frac{x_j}{\pi_j} - X \right)^2 \quad (3.27)$$

$$v(\widehat{X}) = \frac{m}{m-1} \sum_{j=1}^m (k_j - \bar{k})^2, \quad (3.28)$$

où $k_j = w_j x_j$ et $\bar{k} = \frac{1}{m} \sum_{j=1}^m k_j$. Dans le cas de la moyenne nous avons :

$$\widehat{\bar{X}} = \frac{\widehat{X}}{\widehat{M}} = \frac{\sum_{j=1}^m w_j x_j}{\sum_{j=1}^m w_j} = \sum_{j=1}^m \tilde{w}_j x_j, \quad (3.29)$$

où $\tilde{w}_j = w_j / \sum_{k=1}^m w_k$ est le poids normalisé. Comme $\widehat{\bar{X}}$ est un ratio de variables aléatoires, il devient difficile de déduire son erreur quadratique moyenne (EQM) exacte. Il faut alors linéariser cette variable et on obtient :

$$V(\widehat{\bar{X}}) \simeq \frac{1}{M^2} \left[V(\widehat{X}) - 2\widehat{\bar{X}} \cdot COV(\widehat{X}, \widehat{M}) + \widehat{\bar{X}}^2 \cdot V(\widehat{M}) \right] \quad (3.30)$$

$$v(\widehat{\bar{X}}) \simeq \frac{1}{M^2} \left[v(\widehat{X}) - 2\widehat{\bar{X}} \cdot cov(\widehat{X}, \widehat{M}) + \widehat{\bar{X}}^2 \cdot v(\widehat{M}) \right], \quad (3.31)$$

où

$$V(\widehat{M}) = \frac{1}{M} \sum_{j=1}^M \pi_j \left(\frac{1}{\pi_j} - M \right)^2 \quad (3.32)$$

$$v(\widehat{M}) = \frac{m}{m-1} \sum_{j=1}^m (w_j - \bar{w})^2 \quad (3.33)$$

et

$$COV(\widehat{X}, \widehat{M}) = \frac{1}{M} \sum_{j=1}^M \pi_j \left(\frac{x_j}{\pi_j} - X \right) \left(\frac{1}{\pi_j} - M \right) \quad (3.34)$$

$$cov(\widehat{X}, \widehat{M}) = \frac{m}{m-1} \sum_{j=1}^m (k_j - \bar{k}) (w_j - \bar{w}). \quad (3.35)$$

On doit noter que ces deux dernières expressions proviennent de la linéarisation d'un simple ratio $\left(\frac{X}{M}\right)$. Cette formulation est correcte dans le cas de l'estimation de l'erreur quadratique moyenne d'un ratio de variables aléatoires et elle peut être employée pour des plans de sondage plus complexes. Toutefois, il faut s'assurer que les estimateurs de variance des totaux ($v(\widehat{X})$ et $v(\widehat{M})$) et de la covariance ($cov(\widehat{X}, \widehat{M})$) tiennent compte du plan de sondage.

Ici, $V(\widehat{X})$ est l'*EQM* et non la variance car l'estimateur est biaisé. En fait :

$$EQM(\widehat{X}) = E(\widehat{X} - \bar{X})^2 \quad (3.36)$$

Si l'estimateur était non biaisé : $E(\widehat{X}) = \bar{X} \Rightarrow EQM(\widehat{X}) = E(\widehat{X} - E(\widehat{X}))^2 = V(\widehat{X})$. L'estimateur de la moyenne est toutefois convergent. Lorsque la taille de l'échantillon devient suffisamment grande, son biais ($E(\widehat{X}) - \bar{X}$) devient négligeable.

3.1.3 Échantillonnage stratifié

Lorsque l'on désire une meilleure couverture de la population, il est possible d'avoir recours à la stratification. La stratification consiste à subdiviser la population en L sous-ensembles (e.g. zone urbaine et zone rurale) exclusifs. Un échantillon aléatoire sera tiré dans chacun de ces sous-ensembles et L sondages indépendants seront réalisés. Dans le cas de la population, on a :

- M_h , le nombre d'unités échantillonnales dans la strate h ($h = 1, \dots, L$).
- X_{hj} , la mesure de la variable pour l'unité j de la strate h (population).
- $W_h = \frac{M_h}{M}$, taille relative de la strate h .
- $\bar{X}_h = \frac{\sum_{j=1}^{M_h} X_{hj}}{M_h}$, la moyenne de la strate h .
- $S_{xh}^2 = \frac{\sum_{j=1}^{M_h} (X_{hj} - \bar{X}_h)^2}{M_h - 1}$, la variance de la strate h .

On peut étendre les définitions pour le cas de l'échantillon :

- m_h , le nombre d'unités échantillonnées dans la strate h ($h = 1, \dots, L$).
- x_{hj} , la mesure de la variable pour l'unité j de la strate h (échantillon).
- $W_h = \frac{M_h}{M}$, taille relative de la strate h .
- $\bar{x}_h = \frac{\sum_{j=1}^{m_h} x_{hj}}{m_h}$, la moyenne échantillonnale de la strate h .
- $s_{xh}^2 = \frac{\sum_{j=1}^{m_h} (x_{hj} - \bar{x}_h)^2}{m_h - 1}$, la variance échantillonnale de la strate h .

Comme il y a un EAS dans chaque strate, chaque unité échantillonnale de la même strate a un poids identique égal à w_h . La notion de poids échantillonnaux vue précédemment peut tout aussi bien s'appliquer dans le cadre d'un plan stratifié.

Il est possible de démontrer statistiquement que la stratification contribue à améliorer la précision des estimateurs, et principalement en présence d'une forte hétérogénéité inter-strate et d'une forte homogénéité intra-strate quant à la variable d'intérêt (X). Nous

exposons la preuve de ce résultat à l'annexe C, bien qu'on la retrouve dans les références [ASSELIN (1984) et MORIN (1993)] pour la comparaison de l'échantillonnage aléatoire simple et l'échantillonnage aléatoire simple stratifié. Intuitivement, si on désire estimer une variable d'intérêt (et sa précision) auprès des travailleurs du Burkina Faso, par exemple le nombre de semaines de travail rémunérées, il sera alors préférable d'avoir recours au sondage stratifié où chacun des regroupements de travailleurs représentera une strate. Effectivement, les travailleurs d'un même groupe seront davantage semblables (homogénéité intra-strate) par rapport à la variable d'intérêt. Le fait d'interroger plusieurs travailleurs appartenant au même groupe n'apportera pas beaucoup d'information additionnelle quant au nombre de semaines rémunérées. Afin d'assurer une meilleure couverture du Burkina Faso, nous pourrions sélectionner un petit nombre de travailleurs dans chacun des regroupements (disons 1000 travailleurs en tout). La précision de l'estimation du nombre de semaines sera ainsi accrue. En contrepartie, si nous nous contentons de sélectionner plusieurs membres (disons 1000 travailleurs) dans la population à partir d'un plan aléatoire simple (donc sans stratification), l'estimation risque d'être fort mauvaise si ce dernier échantillon est peu représentatif de la population (e.g. il est possible que l'on échantillonne une trop grande proportion de salariés privés par rapport à leur proportion réelle dans la population du Burkina Faso).

Statistiquement, les estimateurs pour un total se présentent ainsi en présence de la stratification :

$$X_{tot} = \sum_{h=1}^L X_{tot(h)} = \sum_{h=1}^L \sum_{j=1}^M X_{hj} \quad (3.37)$$

$$\begin{aligned} \hat{X} &= \sum_{h=1}^L \hat{X}_{tot(h)} = \sum_{h=1}^L \sum_{j=1}^m \frac{M_h}{m_h} x_{hj} \\ &= \sum_{h=1}^L \sum_{j=1}^m w_h x_{hj} \end{aligned} \quad (3.38)$$

$$V(\hat{X}) = \sum_{h=1}^L V(\hat{X}_{tot(h)}) \quad (3.39)$$

$$v(\hat{X}) = \sum_{h=1}^L v(\hat{X}_{tot(h)}) \quad (3.40)$$

Les estimations pour le nombre d'unités dans la population (\widehat{M}) et le ratio de variables aléatoires (\widehat{X}/\widehat{M}) se font comme à la sous-section précédente (équations 3.29 à 3.35).

3.1.4 Échantillonnage à plusieurs degrés

Maintenant que nous avons défini les plans de sondage simples, nous pourrions passer aux plans plus complexes qui ne sont que des combinaisons des sondages simples.

Un problème majeur souvent rencontré lors d'une enquête par sondage est l'inexistence de bases de sondage. Cette situation est courante dans les PED. Il faudrait d'abord recenser la population afin de dresser la liste de toutes les unités statistiques et l'utiliser comme base de sondage. Construire une base de sondage de cette façon est une opération extrêmement coûteuse. Nous aimerions pouvoir sélectionner les unités échantillonnales (e.g. zones de dénombrement et ménages) de façon à ce que l'échantillon soit concentré en un nombre limité de zones tout en demeurant représentatif de la population. La technique utilisée est l'échantillonnage à plusieurs degrés.

Tout d'abord, il faut grouper les unités statistiques (ménages) en plus grosses unités (zones de dénombrement) appelées unités primaires d'échantillonnage (UPE). Par la suite, il faut sélectionner quelques-unes de ces UPE afin de les recenser et d'y dresser la liste des unités secondaires d'échantillonnage (USE). Finalement, nous devons sélectionner un certain nombre d'unités secondaires (ménages) dans chacune des unités primaires (zones) sélectionnées. Ainsi, l'échantillon de ménages est concentré dans un nombre restreint de zones. Les déplacements lors de l'enquête seront réduits, ce qui diminuera le coût et la durée de l'enquête.

La population considérée sera composée de D zones (UPE). Chacune des zones, $c = 1, \dots, D$, regroupe M_c ménages (USE, ou unités statistiques). Nous sélectionnerons d'abord d zones, et ensuite, nous sélectionnerons m_c ménages, $j = 1, \dots, m_c$, dans chacune des d zones tirées. Ici, nous ne traiterons que de l'échantillonnage à deux degrés, quoiqu'il serait toujours possible de recourir à un plan de sondage comprenant plusieurs degrés (UPE, USE, UTE, ..., unités statistiques).

Soit $M = \sum_{c=1}^D M_c$, le nombre total de ménages dans la population. Nous considérerons un échantillonnage avec remise au premier degré (sélection des UPE). La probabilité que le $j^{\text{ième}}$ ménage de la $c^{\text{ième}}$ zone soit sélectionné dans un tirage à deux degrés est $p_{j|c} \times p_c$. Nous pourrions alors procéder comme à la section précédente. Soit $q_c = (dp_c)^{-1}$, l'inverse de la probabilité que la zone c soit échantillonnée au premier degré (c'est le poids de la zone). Soit $q_{j|c} = (m_c p_{j|c})^{-1}$, l'inverse de la probabilité que le ménage j soit tiré parmi les M_c ménages de la zone c (c'est le poids du ménage dans la zone). Alors $w_{cj} = q_c \times q_{j|c}$ constitue le poids, dans la population, du $j^{\text{ième}}$ ménage de la $c^{\text{ième}}$ zone et $w_c = \sum_{j=1}^{m_c} w_{cj}$. Encore une fois, il s'agit d'estimer X et M dans un plan de sondage à deux degrés. Les estimateurs seront toujours \widehat{X} et \widehat{M} :

$$\widehat{X} = \sum_{c=1}^d \sum_{j=1}^{m_c} w_{cj} x_{cj} \quad (3.41)$$

$$\widehat{M} = \sum_{c=1}^d \sum_{j=1}^{m_c} w_{cj}. \quad (3.42)$$

Étant donné que l'on pose l'hypothèse que le tirage des unités primaires a été fait avec remise, l'estimation de la variance pour \widehat{X} s'obtient facilement à partir des poids échantillonnaires ainsi que des estimations du total propres au premier degré (i.e. à partir des w_c et des \widehat{X}_c). Il est à noter que chacun des produits $w_c \widehat{X}_c$ constitue une estimation du total \widehat{X} de la population et que le plan de sondage au deuxième degré n'intervient pas dans l'évaluation de la variance lorsque le tirage au premier degré est supposé avec remise.⁶ Nous estimons la variance comme suit :

$$v(\widehat{M}) = \frac{d}{d-1} \sum_{c=1}^d (w_c - \bar{w})^2 \quad (3.43)$$

$$v(\widehat{X}) = \frac{d}{d-1} \sum_{c=1}^d (k_c - \bar{k})^2 \quad (3.44)$$

$$\text{cov}(\widehat{M}, \widehat{X}) = \frac{d}{d-1} \sum_{c=1}^d (w_c - \bar{w}) (k_c - \bar{k}), \quad (3.45)$$

⁶ Voir la dérivation de l'estimateur de la variance pour un total à l'annexe C.

où $k_c = w_c \widehat{X}_c$ et $\bar{k} = \frac{1}{d} \sum_{c=1}^d k_c$. L'estimation de l'erreur quadratique moyenne du quotient \widehat{X}/\widehat{M} s'obtient aussi facilement à l'aide de l'équation 3.31.

Un cas particulier Considérant le cas où les UPE ont la même probabilité de sélection (d/D , EAS au premier degré), et que dans chacune d'entre elles on tire m USE (plan aléatoire simple au deuxième degré) parmi les M (même nombre d'USE dans chaque UPE), alors la statistique devient :

$$\bar{x}' = \frac{1}{dm} \sum_{c=1}^d \sum_{j=1}^m x_{cj} \quad (3.46)$$

et l'estimateur de la variance dans le cas où les UPE sont tirées avec remise est :

$$v(\bar{x}') = \frac{1}{d(d-1)} \sum_{c=1}^d (\bar{x}_c - \bar{x}')^2 = \frac{1}{d} \widetilde{s}^2, \quad (3.47)$$

où $\bar{x}_c = \frac{1}{m} \sum_{j=1}^m x_{cj}$.

D'un autre côté, si on ignore la structure à deux degrés pour ne considérer qu'un plan aléatoire simple construit à partir d'un échantillon de $d \cdot m$ ménages, nous obtenons l'estimateur naïf⁷ suivant :

$$\bar{x}_{EAS} = \frac{1}{dm} \sum_{c=1}^d \sum_{j=1}^m x_{cj}. \quad (3.48)$$

L'estimateur de la variance est alors :

$$v(\bar{x}_{EAS}) = \frac{1}{dm(dm-1)} \sum_{c=1}^d \sum_{j=1}^m (x_{cj} - \bar{x}_{EAS})^2 = \frac{1}{dm} \widehat{s}^2. \quad (3.49)$$

⁷ On dit qu'il est naïf parce qu'il suppose que les unités statistiques sont obtenues à l'aide d'un échantillonnage aléatoire simple (EAS).

Il est possible de démontrer⁸ que la variance $v(\bar{x}')$ peut être estimée de la façon suivante :

$$v(\bar{x}') \simeq v(\bar{x}_{EAS}) [1 + (m - 1)\hat{\rho}], \quad (3.50)$$

où $\hat{\rho}$ est le *coefficient de corrélation intra-grappe*.⁹ Ce coefficient exprime le degré de ressemblance entre les éléments d'une même unité primaire. Il est défini comme suit :

$$\hat{\rho} = 1 - \frac{s_j^2}{s^2} = 1 - \frac{[d(d-1)]^{-1} \sum_{c=1}^d \sum_{j=1}^m (x_{cj} - \bar{x}_c)^2}{(dm-1)^{-1} \sum_{c=1}^d \sum_{j=1}^m (x_{cj} - \bar{x}')^2}. \quad (3.51)$$

Lorsque les éléments d'une UPE sont aussi dissemblables que les éléments de la population ($s_j^2 = s^2$, $\rho \rightarrow 0$), l'estimateur de variance obtenu avec le plan à deux degrés a sensiblement la même valeur que l'estimateur de variance obtenu avec le plan aléatoire simple. Par contre, lorsque $s_j^2 = 0$ et $\rho \rightarrow 1$ (tous les éléments de la même unité primaire ont la même caractéristique, x_{cj}) :

$$v(\bar{x}') = m \times v(\bar{x}_{EAS}) = \frac{\hat{s}^2}{d}. \quad (3.52)$$

Le fait de mesurer m_c valeurs identiques dans l'UPE n'apporte aucune information supplémentaire sur les unités échantillonnées. La précision dépend plutôt du nombre d'UPE dans l'échantillon et non de la quantité d'USE dans chacune des UPE. La plupart du temps, les individus d'une même classe ou d'une même région ou d'une même grappe ont des caractéristiques assez semblables, ce qui diminue la valeur du coefficient de corrélation intra-grappe. Pour une taille d'échantillon fixe, il devient avantageux d'augmenter le nombre d'unités primaires et de diminuer le nombre d'unités secondaires dans chacune des unités

⁸La preuve est présentée dans COCHRAN (1977).

⁹Ici, le terme «grappe» désigne l'unité primaire d'échantillonnage.

primaires. Il est bien entendu que le plan de sondage à deux degrés est plus précis si $\hat{\rho}$ tend à être nul ($v(\bar{x}') \rightarrow v(\bar{x}_{EAS})$). Le plan de sondage est plus précis si les unités échantillonnées dans une UPE sont hétérogènes entre elles. Chacune des UPE représente alors bien l'ensemble de la population. Notons que nous avons conclu que la stratification était efficace lorsque les strates étaient hétérogènes entre elles et lorsque que les éléments de la strate étaient homogènes. Ces nuances proviennent en fait de la structure du sondage. La stratification assure une bonne couverture de la population alors que ce n'est pas le cas avec l'échantillonnage à plusieurs degrés dû au fait que l'on ne sélectionne qu'un petit nombre de grappes. Contrairement à la stratification, l'échantillonnage à plusieurs degrés aura donc tendance à réduire la précision des estimations.

Introduisons maintenant la stratification dans le cas d'un plan à deux degrés. Supposons que nous ayons deux strates ($L = 2$). La première strate est la région urbaine et la seconde est la région rurale. La stratification nous permet d'avoir une meilleure couverture de la population :

$$\hat{X} = \sum_{h=1}^L \sum_{c=1}^d \sum_{j=1}^m w_{hcj} x_{hcj} \quad (3.53)$$

$$\hat{M} = \sum_{h=1}^L \sum_{c=1}^d \sum_{j=1}^m w_{hcj}, \quad (3.54)$$

où h représente la strate. L'estimation de la variance s'obtient à partir des formules suivantes :

$$v(\hat{M}) = \sum_{h=1}^L \left(\frac{d_h}{d_h-1} \sum_{c=1}^{d_h} (w_{hc} - \bar{w}_h)^2 \right) \quad (3.55)$$

$$v(\hat{X}) = \sum_{h=1}^L \left(\frac{d_h}{d_h-1} \sum_{c=1}^{d_h} (k_{hc} - \bar{k}_h)^2 \right) \quad (3.56)$$

$$cov(\hat{M}, \hat{X}) = \sum_{h=1}^L \left(\frac{d_h}{d_h-1} \sum_{c=1}^{d_h} (w_{hc} - \bar{w}_h) (k_{hc} - \bar{k}_h) \right), \quad (3.57)$$

où $k_{hc} = w_{hc} \widehat{X}_{hc}$ et $\bar{k}_h = \frac{1}{d_h} \sum_{c=1}^{d_h} k_{hc}$. Comme chacune des strates est indépendante des autres par définition, on somme les L variances calculées afin de retrouver la variance des totaux.

De même, pour un ratio :

$$\widehat{\bar{X}} = \frac{\sum_{h=1}^L \widehat{M}_h \bar{x}_{hw}}{\widehat{M}}, \quad (3.58)$$

où

$$\begin{aligned} \widehat{\bar{X}} &= \frac{\sum_{h=1}^L \sum_{c=1}^{d_h} w_{hc} \bar{x}_{hc}}{\sum_{h=1}^L \sum_{c=1}^{d_h} w_{hc}} \\ &= \frac{\sum_{h=1}^L \sum_{c=1}^{d_h} \sum_{j=1}^{m_c} w_{hcj} x_{hcj}}{\sum_{h=1}^L \sum_{c=1}^{d_h} \sum_{j=1}^{m_c} w_{hcj}} = \frac{\widehat{X}_{tot}}{\widehat{M}}. \end{aligned} \quad (3.59)$$

L'estimation de l'erreur quadratique moyenne pour ce quotient se fait de façon similaire à partir de $v(\widehat{M})$, $v(\widehat{X})$, $cov(\widehat{X}, \widehat{M})$ et de l'équation 3.31.

Nous avons mentionné plus tôt que la stratification était profitable en terme de précision lorsque les unités à l'intérieur des strates étaient fortement homogènes. Parallèlement, le fait de regrouper les unités statistiques en plus grosses unités (grappes) accroîtra l'homogénéité à l'intérieur des strates du fait que les grappes diffèrent souvent beaucoup moins entre elles que les unités statistiques. Conséquemment, la stratification sera d'autant plus efficace lorsque nous sommes en présence d'un plan à deux ou plusieurs degrés.

3.1.5 L'effet de plan

L'effet de plan («design effect»), dénoté par DF , est un scalaire qui se définit comme le ratio de la variance obtenue alors que l'on prend en compte la structure du sondage (variance réelle) sur la variance obtenue à partir du plan aléatoire simple, pour un échantillon de même taille. Mathématiquement, l'estimateur de l'effet de plan s'écrit :

$$\widehat{DF} = \frac{v_{réelle}(\hat{\theta})}{v_{EAS}(\hat{\theta})}. \quad (3.60)$$

L'effet de plan permet de juger de l'importance du plan de sondage lors de l'estimation de la variance de l'estimateur θ . Plus l'effet de plan sera différent de l'unité, plus il sera erroné de ne pas respecter le plan de sondage employé sur le terrain par les enquêteurs. Un effet de plan inférieur à l'unité signifie que l'on surestime la vraie variance en posant l'hypothèse d'un EAS, alors qu'un autre supérieur à l'unité signalera une sous-estimation de la vraie variance.

3.1.6 Les intervalles de confiance

Une estimation d'un paramètre sous forme d'intervalle sera parfois préférée à une estimation ponctuelle associée à une mesure de précision. L'intervalle de confiance fournit un ensemble de valeurs probables pour un paramètre d'intérêt. En présence de grands échantillons, l'approximation normale s'avère très satisfaisante quand il s'agit d'estimer la distribution d'un total ou d'un ratio. Ainsi, l'intervalle de confiance de niveau $1 - \alpha$ (α compris entre 0 et 1) pour le paramètre θ aura la forme suivante :

$$I.C. : \left[\hat{\theta} \pm z_{(1-\frac{\alpha}{2})} \sqrt{v(\hat{\theta})} \right], \quad (3.61)$$

où $z_{(1-\frac{\alpha}{2})}$ est le $(1 - \frac{\alpha}{2})^{\text{ième}}$ quantile de la loi normale centrée réduite ($N(0, 1)$).

3.2 Estimation des indices de pauvreté et d'inégalité

Le plan de sondage employé pour les fins de l'Enquête Prioritaire tenue au Burkina Faso était stratifié et comportait deux degrés d'échantillonnage avec probabilités de sélection inégales au premier degré (tirage des unités primaires d'échantillonnage, soit les ZD) et probabilités de sélection égales au deuxième degré (tirage des ménages dans chacune des ZD sélectionnées au degré 1). Le pays fut subdivisé en sept strates ($L = 7$) et en 434 zones de dénombrement ($D = 434$). Les strates n'ont pas nécessairement un nombre égal de zones de dénombrement. Le fichier de données fourni par l'INSD comporte une multitude de renseignements sur un échantillon de $S = 8642$ ménages. Le fichier initial contenait les poids échantillonnaires des ménages sélectionnés (variable w) de même que la taille des ménages en individus (variable T). L'indicateur du niveau de vie était le montant des dépenses réelles per capita (Y). Nous savons également qu'un échantillon de d_h zones fut tiré dans chacune des strates ($h = 1, 2, \dots, 7$) et que $m_c = 20$ ménages furent tirés dans chacune des zones.

3.2.1 Les dépenses et la population

L'estimateur des dépenses totales est donné par :

$$\widehat{Y} = \sum_{h=1}^L \sum_{c=1}^{d_h} \sum_{j=1}^{m_c} w_{hcj} t_{hcj} y_{hcj} \quad (3.62)$$

et l'estimateur de sa variance :

$$v(\widehat{Y}) = \sum_{h=1}^L \frac{d_h}{d_h - 1} \sum_{c=1}^{d_h} (z_c - \bar{z})^2, \quad (3.63)$$

où $z_c = \sum_{j=1}^{m_c} w_{hcj} t_{hcj} y_{hcj}$ et $\bar{z} = \frac{1}{d_h} \sum_{c=1}^{d_h} z_c$.

L'estimateur du nombre total d'individus est donné par :

$$\hat{N} = \sum_{h=1}^L \sum_{c=1}^{d_h} \sum_{j=1}^{m_c} w_{hcj} t_{hcj} \quad (3.64)$$

et l'estimateur de la variance utilisé est :¹⁰

$$v(\hat{N}) = \sum_{h=1}^L \frac{d_h}{d_h - 1} \sum_{c=1}^{d_h} (u_c - \bar{u})^2, \quad (3.65)$$

où $u_c = \sum_{j=1}^{m_c} w_{hcj} t_{hcj}$ et $\bar{u} = \frac{1}{d_h} \sum_{c=1}^{d_h} u_c$.

3.2.2 Le niveau moyen des dépenses réelles per capita

Désormais, pour alléger l'écriture, nous écrivons $\sum_S w_s t_s y_s$. Cette notation équivaut à $\sum_{h=1}^L \sum_{c=1}^{d_h} \sum_{j=1}^{m_c} w_{hcj} t_{hcj} y_{hcj}$.

L'estimateur des dépenses réelles per capita est donné par :

$$\hat{Y} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_S w_s t_s y_s}{\sum_S w_s t_s}. \quad (3.66)$$

¹⁰Nous avons utilisé le progiciel IMPS pour les calculs de l'erreur échantillonnale. IMPS est un acronyme pour «Integrated Microcomputer Processing System». Ce progiciel permet d'estimer l'erreur échantillonnale de totaux et de ratios dans le cadre de plans de sondage complexes.

L'estimateur de son *EQM* est donné par :

$$v(\widehat{Y}) = \frac{1}{\widehat{N}^2} \left[v(\widehat{Y}) - 2\widehat{Y} \cdot \text{cov}(\widehat{Y}, \widehat{N}) + \widehat{Y}^2 \cdot v(\widehat{N}) \right], \quad (3.67)$$

où

$$\text{cov}(\widehat{Y}, \widehat{N}) = \frac{L}{\sum_{h=1}^L d_h - 1} \frac{d_h}{\sum_{c=1}^{d_h} (z_c - \bar{z})(u_c - \bar{u})}. \quad (3.68)$$

3.2.3 Les indices de pauvreté de la classe FGT

L'estimateur des indices FGT est donné par :

$$\widehat{P}_\alpha = \frac{\widehat{Q}}{\widehat{N}} = \frac{\sum w_s t_s \left(\frac{z - y_s}{z} \right)^\alpha \mathbb{1}(y_s \leq z)}{\sum w_s t_s} = \frac{\sum w_s t_s q_s}{\sum w_s t_s}, \quad (3.69)$$

où q_s représente une mesure de pauvreté individuelle.

L'estimateur de l'*EQM* qu'utilise IMPS est :

$$v(\widehat{P}_\alpha) = \frac{1}{\widehat{N}^2} \left[v(\widehat{Q}) - 2\widehat{P}_\alpha \cdot \text{cov}(\widehat{Q}, \widehat{N}) + \widehat{P}_\alpha^2 \cdot v(\widehat{N}) \right] \quad (3.70)$$

$$\text{cov}(\widehat{Q}, \widehat{N}) = \frac{L}{\sum_{h=1}^L d_h - 1} \frac{d_h}{\sum_{c=1}^{d_h} (r_c - \bar{r})(u_c - \bar{u})}, \quad (3.71)$$

où $r_c = \sum_{j=1}^{m_c} w_{hcj} t_{hcj} q_{hcj}$ et $\bar{r} = \frac{1}{d_h} \sum_{c=1}^{d_h} r_c$.

3.2.4 L'indice d'inégalité d'Atkinson

L'estimateur de A_ϵ s'écrit

$$\widehat{A}_\epsilon = 1 - \left[\left(\frac{1}{\sum_S w_s t_s} \right) \frac{\left(\sum_S w_s t_s \right)^{1-\epsilon}}{\left(\sum_S w_s t_s y_s \right)^{1-\epsilon}} \left(\sum_S w_s t_s (y_s)^{1-\epsilon} \right) \right]^{\frac{1}{1-\epsilon}}. \quad (3.72)$$

Après plusieurs simplifications :

$$\widehat{A}_\epsilon = 1 - \left[\left(\frac{1}{\sum_S w_s t_s y_s} \right) \frac{\left(\sum_S w_s t_s (y_s)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}}{\left(\sum_S w_s t_s \right)^{\frac{1}{1-\epsilon}}} \right] \quad (3.73)$$

$$\begin{aligned} &= 1 - \left[\left(\frac{1}{\widehat{Y}} \right) \left(\frac{\widehat{Z}^a}{\widehat{X}^{a-1}} \right) \right] \\ &= 1 - \widehat{\theta}, \end{aligned} \quad (3.74)$$

en posant $a = \frac{1}{1-\epsilon}$.

Si nous désirons estimer l'erreur quadratique moyenne, nous devons procéder à une linéarisation de l'expression entre crochets. Voici un bref aperçu de la linéarisation.

Prenons d'abord le logarithme de cette expression :

$$\ln \widehat{\theta} = a \ln \widehat{Z} - (a - 1) \ln \widehat{X} - \ln \widehat{Y}. \quad (3.75)$$

Utilisons ensuite le résultat suivant :

$$v(\ln \widehat{\theta}) \simeq \left[\frac{\partial \ln \theta}{\partial \theta} \Big|_{\theta=\widehat{\theta}} \right]^2 v(\widehat{\theta}) \quad (3.76)$$

ce qui implique que

$$v(\hat{\theta}) = \left[\frac{\partial \ln \theta}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right]^{-2} v(\ln \hat{\theta}). \quad (3.77)$$

Par la suite, nous approximations $\ln \hat{\theta} - \ln \theta^{11}$ selon la méthode de TAYLOR :

$$\begin{aligned} \ln \hat{\theta} - \ln \theta &\simeq \frac{a}{Z}(\hat{Z} - Z) - \frac{(a-1)}{\bar{X}}(\hat{X} - X) - \frac{(\hat{Y}-Y)}{\bar{Y}} \\ &= \frac{a\hat{Z}}{Z} - (a-1)\frac{\hat{X}}{\bar{X}} - \frac{\hat{Y}}{\bar{Y}} \\ &\simeq \sum_S w_s t_s \left[\frac{a \cdot y_s^{1-\epsilon}}{Z} - \frac{(a-1)}{\bar{X}} - \frac{y_s}{\bar{Y}} \right] \\ &= \sum_S w_s t_s [k_s]. \end{aligned} \quad (3.78)$$

Comme $\ln \theta$ est constant, $v(\ln \hat{\theta} - \ln \theta) = v(\ln \hat{\theta})$. On estime donc la variance de $\ln \hat{\theta}$ à partir d'une variable auxiliaire, k . On aura tout simplement à estimer la variance de la somme des k_s . Finalement, nous isolons la variance estimée de $\hat{\theta}$ comme suit :

$$v(\hat{\theta}) = \frac{v(\ln \hat{\theta})}{\left[\frac{\partial \ln \theta}{\partial \theta} \Big|_{\theta=\hat{\theta}} \right]^2} = \frac{v(\ln \hat{\theta})}{\left[\frac{1}{\hat{\theta}} \right]^2} = \hat{\theta}^2 v(\ln \hat{\theta}) = v(A_\epsilon), \quad (3.79)$$

ce qu'il fallait démontrer.¹²

¹¹ Il est à noter que θ , non indicé, est un terme constant.

¹² La dérivation de ce résultat est donnée à l'annexe C.

3.2.5 La dominance stochastique

Dans le cas de la dominance stochastique, nous avons retenu que les points d'une courbe de dominance stochastique étaient calculés à l'aide des indices FGT. Voici un bref rappel de la théorie :

- 1) *Courbe d'incidence de la pauvreté* : $F(z) = P_0(z)$. Or, P_0 est un ratio de variables aléatoires ;
- 2) *Courbe du déficit de la pauvreté* : $D(z) = z \cdot P_1(z)$, où z est considéré constant et P_1 est un ratio de variables aléatoires ;
- 3) *Courbe de l'intensité de la pauvreté* : $S(z) = \frac{z^2}{2} \cdot P_2(z)$, où z est constant et P_2 est aussi un ratio de variables aléatoires.

Les courbes de dominance stochastique sont utilisées dans le but de rechercher une comparaison normativement robuste entre deux situations. Il serait pertinent de pouvoir vérifier statistiquement si une courbe (situation A) est significativement supérieure à une autre (situation B) pour un seuil de pauvreté donné (z^*). Nous pourrions ensuite tester la différence entre ces deux courbes sur l'intervalle de seuils de pauvreté admissibles.

La méthode consiste à calculer un intervalle de confiance de niveau $1-\alpha$ pour la différence entre deux courbes, ou plutôt entre deux ratios. Pour $S(z)$, par exemple :

$$\widehat{S}_A(z^*) = \frac{z^2}{2} \widehat{P}_{2A}(z^*) = \frac{(z^*)^2}{2} \cdot \frac{\sum_{S(A)} w_s t_s (q_s^{(2)})}{\sum_{S(A)} w_s t_s} = \frac{(z^*)^2}{2} \cdot \widehat{Q}_A \quad (3.80)$$

$$\widehat{S}_B(z^*) = \frac{z^2}{2} \widehat{P}_{2B}(z^*) = \frac{(z^*)^2}{2} \cdot \frac{\sum_{S(B)} w_s t_s (q_s^{(2)})}{\sum_{S(B)} w_s t_s} = \frac{(z^*)^2}{2} \cdot \widehat{Q}_B. \quad (3.81)$$

Nous voulons ensuite estimer l'*EQM* de la différence :

$$\Delta \widehat{S}(z^*) = \widehat{S}_A(z^*) - \widehat{S}_B(z^*) = \frac{(z^*)^2}{2} \cdot \left[\frac{\sum_{S(A)} w_s t_s (q_s^{(2)})}{\sum_{S(A)} w_s t_s} - \frac{\sum_{S(B)} w_s t_s (q_s^{(2)})}{\sum_{S(B)} w_s t_s} \right]. \quad (3.82)$$

L'estimateur utilisé par IMPS s'écrit :

$$eqm(\Delta \widehat{S}(z)) \simeq \sum_{h=1}^L \frac{d_h}{d_h - 1} \left(1 - \frac{d_h}{D_h}\right) \sum_{c=1}^{d_h} (\delta_{hc} - \bar{\delta}_{h..})^2, \quad (3.83)$$

où d_h est le nombre d'UPE de la strate h dans l'échantillon et D_h est le nombre total d'UPE dans la strate h . De plus :

$$\delta_{hc} = \sum_{j=1}^{m_{hc}} w_{hcj} (\Lambda_{hcj(A)} - \Lambda_{hcj(B)}), \quad (3.84)$$

$$\Lambda_{hcj(l)} = \frac{z^2/2}{\widehat{N}(l)} \cdot [t_{hcj(l)} q_{hcj(l)} - \widehat{S}(l) t_{hcj(l)}], \quad (3.85)$$

$$\bar{\delta}_{h..} = \frac{1}{d_h} \sum_{c=1}^{d_h} \delta_{hc}, \quad (3.86)$$

où $l = A, B$.

Dans ce cas-ci, m_{hc} est le nombre d'USE sélectionnées dans la $c^{i\text{ème}}$ UPE de la $h^{i\text{ème}}$ strate.

À partir de la racine carrée de l'erreur quadratique moyenne estimée pour la différence, il sera possible de calculer un intervalle de confiance. Si l'intervalle $\widehat{S}_A(z) - \widehat{S}_B(z)$ ne contient que des valeurs positives pour tous les seuils admissibles $z \in (0, z_{\max})$, alors la Condition de dominance du troisième ordre sera respectée et nous aurons une comparaison normativement et statistiquement robuste.

Chapitre 4

Données de l'Enquête Prioritaire

Dans le cadre de son «Programme d'Ajustement Structurel» (PAS), le gouvernement du Burkina Faso avait besoin de données exhaustives, fiables et à jour sur les différents secteurs de l'économie, dont le niveau de vie des ménages. On décida donc de mener pour la première fois une enquête sur les conditions de vie des ménages afin de produire des indicateurs socio-économiques sur l'ensemble des ménages burkinabés et d'identifier les caractéristiques et la localisation des sous-groupes socio-économiques les plus vulnérables. Le programme d'évaluation de la pauvreté au Burkina Faso comporte quatre phases qui sont :

- la collecte des données sur les conditions de vie des ménages ;
- l'étude du profil de pauvreté des Burkinabés dans le but de faire ressortir l'incidence, la gravité, les tendances et les causes de la pauvreté ;
- la mise en oeuvre d'une stratégie nationale de lutte contre la pauvreté ;
- et la phase d'opérationnalisation de la stratégie nationale.

C'est dans ce contexte que le Centre de recherche pour le développement international (CRDI) invita le CECI¹ et le CRÉFA² à établir conjointement un programme d'appui

¹Centre canadien d'étude et de coopération internationale, organisme privé à but non lucratif.

²Centre de recherche en économie et finance appliquées

technique à l'ensemble des équipes MIMAP afin de venir en aide aux pays ayant moins d'expertise au plan de la recherche en économie (mesure et analyse de la pauvreté et de l'inégalité) et en statistique (méthodologie d'enquêtes auprès des ménages).

Le plan de sondage employé au cours de l'Enquête Prioritaire fut élaboré par l'Institut National de la Statistique et de la Démographie [INSD (1996)]. Le plan de sondage devait tenir compte des trois contraintes suivantes :

- le choix de la taille de l'échantillon ;
- les ressources humaines et matérielles disponibles ;
- et l'utilisation de la base de sondage (listes des unités échantillonales) de l'enquête démographique de 1991.

Le pays a d'abord été découpé en sept strates (variable STRATE) ou en huit strates d'analyse (variable STRATANA).

Identification	STRATE	STRATANA	Nombre de ZD (STRATE)	Nombre d'observations (STRATANA)
Ouest	1	1	42	840
Sud	2	2	37	457
Centre nord	3	3	97	1959
Centre sud	4	4	55	1098
Nord	5	5	66	1290
Sud-Est	2	6	37	280
Autres villes	6	7	40	780
Ouagadougou + Bobo-Dioulasso	7	8	98	1938
		Totaux	434	$n = 8642$

Il y a deux STRATES en ZONE urbaine et cinq en ZONE rurale :

Milieu	STRATE	STRATANA
urbain	6 et 7	7 et 8
rural	1 à 5	1 à 6

Le mode de sondage adopté fut celui d'un sondage stratifié (7 strates) à deux degrés (les zones de dénombrement, ZD, au premier degré, et les ménages au deuxième degré). Les ZD sont celles qui ont été établies au recensement démographique de 1985.

Le tirage de l'échantillon dans la strate 7 «Ouagadougou–Bobo-Dioulasso» s'est fait comme suit :

- 1- tirage des ZD (degré 1) avec probabilités égales et sans remise ;
- 2- tirage de 20 ménages dans les ZD de l'échantillon.

Le tirage de l'échantillon dans la strate 6 «Autres villes» et les strates 1 à 5 «strates rurales» s'est fait comme suit :

- 1- au degré 1, tirage des ZD avec probabilités proportionnelles à la taille sans remise (taille en ménages selon le recensement de 1985) ;
- 2- tirage d'environ 20 ménages par ZD.

Le fichier de 8642 observations obtenu suite à l'Enquête Prioritaire de l'INSD fut découpé en plusieurs domaines d'étude (sous-groupes). Les variables STRATANA et ZONE constituent les premières variables de classification. Il y a également les variables GSE (groupes socio-économiques), TYPE (type de ménage), SEXE (sexe du chef de ménage) et TMEN (taille du ménage). Pour les fins de notre étude empirique, nous avons retenu les variables de classification suivantes :

1- groupes socio-économiques (GSE) ;

GSE	Identification	Nombre d'observations
1	salariés du secteur public	675
2	salariés du secteur privé	482
3	artisans et commerçants	1026
4	autres actifs	104
5	agriculteurs rentiers	486
6	agriculteurs vivriers	5154
7	inactifs	715
	Total	$n = 8642$

2- strates d'analyse (STRATANA) ;

3- zones géographiques (ZONE).

L'ensemble des variables utilisées pour nos estimations est le suivant :

STRATE	: strates échantillonnales ;
ZD	: unités primaires d'échantillonnage (zones de dénombrement) ;
STRATANA	: strates d'analyse (classification ou domaine d'étude) ;
GSE	: groupes socio-économiques (classification ou domaine d'étude) ;
ZONE	: zones géographiques (classification ou domaine d'étude) ;
POIDS1 (W)	: poids échantillonnaires des ménages ;
TMEN (T)	: taille des ménages (nombre d'individus) ;
NVIE (Y)	: dépenses réelles per capita (variable d'intérêt).

Pour la saisie des données, l'INSD a utilisé les logiciels ISSA et SPSS. Le fichier initial que nous avons reçu était en format SPSS. Pour nos estimations, nous avons ensuite utilisé les logiciels SPSS (estimation des indicateurs), IMPS (estimation de l'erreur échantillonnaire des indicateurs économiques) ainsi que le logiciel GAUSS de la firme APTECH

inc. (transformations plus complexes du fichier SPSS initial et pour les tests de dominance stochastique).

Les résultats empiriques sont présentés au prochain chapitre. Nous avons estimé les indicateurs suivants ainsi que leur erreur échantillonnale (erreur-type) : le niveau des dépenses réelles totales, la population totale, le niveau des dépenses réelles per capita, les indices FGT, l'indice d'ATKINSON, l'ordonnée des courbes de dominance stochastique (incidence, déficit et intensité de la pauvreté) de même que l'écart entre deux courbes de dominance stochastique dans le cas de la comparaison de deux situations. Nous avons aussi estimé un profil de pauvreté. Finalement, un parallèle avec la théorie statistique présentée au chapitre «**Méthodologie**» permettra de mettre en relief l'**impact du plan de sondage** sur la qualité des estimations de l'erreur, via l'effet de plan.

Comme il y avait des observations manquantes pour le niveau de vie (variable *NVIE*) de 14 ménages et que le logiciel IMPS ne les acceptent pas, nous avons choisi d'imputer des valeurs pour cette variable. La méthode d'imputation utilisée est fort simple. Nous avons estimé le niveau de vie moyen pour chacune des huit strates d'analyse, pour ensuite remplacer la valeur manquante par le niveau de vie moyen de la strate d'analyse correspondante. Les niveaux de vie moyens ont été calculés à partir des poids ménages³ (variable *POIDS1*). La méthode visant à retirer les observations (réduire l'échantillon à 8628 observations) pour lesquelles le niveau de vie n'est pas observé se défend tout aussi bien et les estimations obtenues seront au moins d'aussi bonne qualité. Le taux de couverture pour la variable *NVIE* est estimé à 0,9975% ou 0,9972% selon que l'on utilise le poids ménage (*POIDS1*) ou le poids individu (*TMEN * POIDS1*). Le taux de couverture s'estime ainsi :

$$TX = \frac{\sum_{i=1}^{n-k} W_i}{\sum_{i=1}^n W_i}, \quad (4.87)$$

³Ces poids furent calculés par l'INSD à l'aide des données du recensement de 1985 et de l'enquête démographique de 1991.

où W_i est le poids retenu, n est la taille de l'échantillon initial ($n = 8642$) et k est le nombre d'observations pour lesquelles la variable $NVIE$ n'est pas observée ($k = 14$). La méthode d'imputation utilisée dans notre analyse permet d'expliquer les différences minimales entre les résultats obtenus et les résultats officiels. Étant donné le très faible taux de non réponse (inférieur à 0,03%) pour le niveau de vie réel, les deux approches (imputation de valeurs us retraits des observations) produiront sensiblement les mêmes résultats. À titre d'exemple, nous obtenons un niveau de vie moyen de 72 395 f.CFA par rapport à 72 783 f. CFA pour le Ministère de l'Économie, des Finances et du Plan (1996). Cela constitue une différence d'environ 0,053%, ce qui est peu significatif.

Chapitre 5

Résultats et analyse empirique

Les résultats du mémoire s'ont rapportés aux annexes A et B. L'analyse subséquente fera d'ailleurs référence à ces tableaux et figures. Cependant, les tableaux ne sont pas exhaustifs. Comme il n'était pas possible de regrouper la totalité des résultats sous une forme compacte et conviviale, nous avons décidé de présenter un nombre restreint de résultats.

5.1 Résultats sommaires

5.1.1 Le Burkina Faso

La population totale du Burkina Faso fut estimée à $9\,392\,567 \pm 705\,875$ individus (I.C. 95%) et le nombre moyen d'individus par ménage est de 7,75. On constate que la majorité des individus habitent la zone rurale, soit 83,8% de la population. Ensuite, nous avons remarqué que les strates d'analyse les plus peuplées sont le «Centre sud» et le «Centre nord», regroupant 23 % et 24 % (soit un total de 47%) des Burkinabés. Pareillement, la grande majorité des habitants sont des agriculteurs vivriers, soit 68,1% de la population.

Quant au bien-être des individus, le niveau de vie réel total fut estimé à 683 531 650 000 ± 63M. f. CFA. et le niveau de vie réel per capita se situe aux environs de 72 395 ± 3 604 f. CFA (é.-t. 1 839 f. CFA). Nous observons que le secteur rural représente seulement 36% des dépenses réelles totales. Dans le même ordre d'idée, les strates d'analyse «Ouest», «Centre sud» et «Ouagadougou–Bobo-Dioulasso» comptent pour 62% des dépenses réelles totales alors que les groupes des «salariés publics» et des «agriculteurs vivriers» (ces derniers étant très nombreux) comptent pour 63% des dépenses.

D'autre part, nous avons estimé la proportion d'individus sous le seuil de pauvreté (41 099 f. CFA) à 44,4% (é.-t. 1,6%). De plus, l'écart de pauvreté ainsi que la sévérité de la pauvreté furent estimés à 13,9% (0,7%) et à 6% (0,3%). Certains de ces résultats présentés précédemment se trouvent en annexe. Voir les tableaux A.1 et A.2.

L'étude de la population globale nous permet d'obtenir des estimations pour l'indice d'inégalité d'ATKINSON. Toutefois, ces résultats ont très peu d'intérêt, car il s'agit d'un indice ordinal servant à comparer des domaines d'étude. Nous discuterons davantage de ces estimations dans les prochains paragraphes. Voir les résultats au tableau A.3.

5.1.2 Les zones rurale et urbaine

Tel que présenté dans «Le Profil de Pauvreté au Burkina Faso» [INSD, (1996)] et dans «Le Profil d'inégalité et de pauvreté au Burkina Faso» [WETTA et al., (2000)], nous constatons que la pauvreté est un phénomène plutôt rural tandis que l'inégalité du bien-être en est un urbain. Le niveau de vie réel per capita est estimé à 55 571 f. CFA (é.-t. 1 833 f.) pour la zone rurale et à 159 460 f. CFA (é.-t. 6 721 f.) pour la zone urbaine. Nous observons le même phénomène en comparant les trois indices FGT pour ces deux régions. Les indices FGT propres à la zone rurale ont effectivement des valeurs plus élevées qu'en milieu urbain (annexes A.1 et A.2). Nous remarquons également que les estimations pour les indicateurs

du niveau de vie sont beaucoup plus volatiles dans la zone urbaine (les coefficients de variation des estimateurs, C.V., sont jusqu'à trois fois plus élevés dans cette zone), pour une plus grande dispersion du niveau de vie.

En examinant les courbes d'incidence de la pauvreté pour ces deux zones (annexes B.5 et B.6), on peut conclure qu'il y a Dominance du premier ordre (il y a donc dominance du second et troisième ordre) pour l'intervalle de seuils considéré (15 000 - 80 000 f. CFA). Nous n'avons pas estimé les courbes pour les seuils inférieurs à 15 000 f. afin de réduire le nombre de calculs et de préserver la clarté des images. En fait, les deux courbes deviennent trop rapprochées pour qu'il soit possible de les différencier visuellement ou d'affirmer qu'une est significativement supérieure à une autre. Nous avons alors choisi de tester la robustesse des comparaisons de la pauvreté sur un intervalle de seuils jugés plausibles, soit l'intervalle : [15 000 - 80 000 f. CFA]. En deçà de 15 000 f. (seuils non admissibles), nous considérerons que les conditions de dominance sont respectées, même si en théorie on devrait appliquer les tests de dominance stochastique sur l'intervalle $[0, z]$.

Les indices d'ATKINSON estimés à partir de l'échantillon sont plus élevés pour la zone urbaine. Pour $\varepsilon = 0,25$, nous obtenons un indice rural de 0,067 et un indice urbain de 0,091. Pour $\varepsilon = 10$, les indices sont de 0,826 contre 0,857. Comme le prévoit la théorie économique, nous constatons que l'indice d'inégalité tend à augmenter au fur et à mesure que le paramètre ε augmente. Toutefois, l'écart relatif entre l'inégalité des deux zones géographiques tend à diminuer avec la valeur de ε . Ce résultat n'est pas très surprenant du fait qu'une valeur très élevée pour ce paramètre indique une aversion sociale extrême pour l'inégalité. Les indices convergeront alors très rapidement vers la valeur maximale, soit l'unité, dès qu'il y aura inégalité dans une zone géographique. La dernière remarque s'applique aussi au cas des indices de pauvreté. En effet, l'écart relatif entre les estimations de P_α pour les deux zones géographiques diminue avec l'accroissement du facteur d'aversion pour la pauvreté (α).

Les constats ci-haut s'expliquent entre autres choses par le fait que la taille moyenne des ménages ruraux est supérieure à celle en milieu urbain (8,06 individus contre 6,48) et que

le niveau de vie per capita est très souvent corrélé négativement avec la taille du ménage. Les ménages de grande taille comportent un plus grand nombre d'enfants dont le revenu est négligeable par rapport à leur niveau de consommation. Cela explique partiellement pourquoi la pauvreté est plus élevée en milieu rural. Le niveau de vie en milieu urbain sera davantage inégal du fait que la population y est beaucoup plus diversifiée. On y retrouve des salariés publics et privés, des artisans, des commerçants, des inactifs, etc. Les individus de la zone rurale sont, pour la plupart, des agriculteurs, ce qui contribue à augmenter l'égalité entre les ménages. De plus, le fait que les ménages urbains soient de plus petite taille accentue l'inégalité entre les ménages, et donc l'inégalité entre les individus.¹

5.1.3 Les groupes socio-économiques

Au niveau de l'analyse de la pauvreté, le niveau moyen de vie per capita ainsi que les indices FGT fournissent des mesures cohérentes. Le groupe le plus pauvre semble être celui des agriculteurs vivriers, suivi du groupe des agriculteurs rentiers (provenant principalement de la zone rurale). Les plus nantis semblent être les salariés publics, suivis des salariés privés (provenant largement de la zone urbaine). Le niveau moyen de vie pour les sept groupes se classe ainsi en ordre croissant : agriculteurs vivriers, agriculteurs rentiers, inactifs, autres actifs, artisans, commerçants, salariés privés et salariés publics.

Les courbes d'incidence de la pauvreté permettent d'ordonner partiellement le niveau de pauvreté de quelques groupes. Notons qu'il y a confusion entre les salariés privés et les artisans ainsi qu'entre les deux types d'agriculteurs pour des seuils de pauvreté supérieurs à 33 000 f. CFA. Il semble donc y avoir Dominance du premier ordre au niveau de certains groupes socio-économiques si on considère des seuils de pauvreté assez élevés (33 000 f. CFA et plus). Par contre, pour les seuils inférieurs à 33 000 f. CFA, les écarts entre les courbes d'incidence risquent de ne pas être statistiquement significatifs. Pour plus de rigueur, nous

¹ WETTA C. et al. mentionnent que les ménages de plus petite taille connaîtraient une plus forte inégalité. D'autre part, les ménages de grande taille seraient plus égalitaires quant au bien-être de chacun des membres.

sommes conscients qu'il serait encore nécessaire de tester la signification des écarts au niveau des seuils plus faibles (0 à 33 000 f. CFA). Toutefois, ces seuils de pauvreté sont peu plausibles dans le cas de la population choisie.

D'autre part, les courbes du déficit de la pauvreté pour les agriculteurs vivriers versus les inactifs ne permettent pas d'établir la Dominance du second ordre. Pareillement, la superposition des courbes de l'intensité de la pauvreté pour ces deux groupes ne permet pas de confirmer la Dominance du troisième ordre. Voir les figures B.3 et B.4.

On constate, malheureusement, que les courbes d'incidence de la pauvreté des quatre premiers groupes comportent de nombreux plateaux pour les seuils allant de 15 000 à 30 000 f. CFA. Ce résultat, plutôt surprenant, est sans doute attribuable au manque d'observations (échantillon de petite taille, e.g. 104 ménages pour le GSE «autres actifs») dans certains domaines d'étude, ce qui occasionne les sauts brusques dans les proportions d'individus sous le seuil de pauvreté. Le lissage des courbes dépend également de l'incrément accordé (i.e. un incrément faible) aux valeurs du seuil de pauvreté. Voir la figure B.4.

Au niveau de l'inégalité, les «salariés publics» et les «agriculteurs» semblent avoir un niveau de vie plus homogène que les groupes «autres actifs» et «salariés privés». Bien que le niveau de vie moyen soit assez hétérogène entre les groupes (51 657 f. CFA chez les inactifs contre 252 270 f. CFA chez les salariés publics), il est plutôt difficile de comparer les indices entre les groupes pour toutes valeurs du paramètre. Les classements diffèrent en fonction du choix de ce dernier. Les salariés du secteur public sont souvent plus rapprochés et ont des intérêts communs, ce qui leur permet d'être mieux organisés afin d'obtenir des conditions de vie supérieures et similaires. À titre d'exemple, citons le salaire minimum interprofessionnel garanti (SMIG). D'un autre côté, le niveau de vie des agriculteurs, occupant principalement la zone rurale, sera peu volatile pour les raisons mentionnées auparavant. Par opposition, on peut s'attendre à ce que le groupe «autres actifs» comporte des individus ayant des statuts et des niveaux de vie assez différents. Il n'est donc pas surprenant d'y observer un niveau d'inégalité plus élevé. Pour la plupart des valeurs du paramètre ($0,25 \leq \varepsilon \leq 5$),

le classement en ordre décroissant d'inégalité va comme suit : autres actifs, salariés privés, inactifs, commerçants, artisans, salariés publics, agriculteurs rentiers et agriculteurs vivriers.

5.1.4 Les strates d'analyse

Si on compare le niveau de vie per capita ainsi que les indices P_0 , P_1 et P_2 , on constate que les strates d'analyse les plus pauvres se situent dans le nord («Centre nord» et «Nord») ainsi que dans le sud («Centre sud» et «Sud-Est») de la zone rurale. Les moins pauvres sont encore les strates urbaines («Ouagadougou et Bobo-Dioulasso» ainsi que les «Autres villes»). La taille moyenne des ménages vivant dans le «Centre nord» et le «Centre sud» rural est plus élevée que celle des ménages vivant dans les deux grandes villes (8,54 individus/ménage contre 6,14 individus/ménage), ce qui tend à accroître la pauvreté de même que l'égalité. De plus, la pauvreté plus élevée pour ces régions rurales peut aussi dépendre des conditions climatiques (e.g. sécheresse) et du type d'agriculture pratiquée (e.g. élevage).

À première vue, les courbes d'incidence de la pauvreté (annexes B.3 et B.4) semblent montrer une Dominance du premier ordre au niveau des strates d'analyse de la zone urbaine. Toutefois, les six courbes en zone rurale ne montrent pas un classement clair et non ambigu pour certaines strates d'analyse (e.g. le «Sud-Est» et le «Centre sud»). On n'a qu'à remarquer les nombreux croisements au niveau des seuils faibles (15 000 - 30 000 f. CFA). Se rapporter à la figure B.2.

Le classement des strates d'analyse en matière d'inégalité est plutôt ambigu. Ce dernier dépend du choix du paramètre d'aversion pour l'inégalité. Pour $\varepsilon = 0,5$, nous avons en ordre décroissant d'inégalité : les «Autres villes, Ouagadougou–Bobo-Dioulasso, Sud, Nord, Ouest, Sud-Est, Centre Sud, et Centre nord». On remarque une fois de plus que l'inégalité est supérieure en milieu urbain. En milieu urbain le niveau de vie sera plus égalitaire dans les deux grandes villes, sans doute parce qu'on y retrouve une masse importante de salariés du

secteur public tandis que dans les «Autres villes» on retrouve une économie beaucoup plus diversifiée. De surcroît, les strates d'analyse les plus pauvres (en zone rurale) ont un niveau de vie plus également réparti. Rappelons que les strates rurales regroupent principalement des agriculteurs. Par contre, pour une valeur $\varepsilon = 5$, le classement en ordre décroissant d'inégalité sera : «Nord, Ouagadougou–Bobo-Dioulasso, Autres villes, Ouest, Sud, Centre nord, Sud-Est et Centre Sud». Ce classement s'interprète plus difficilement étant donné la présence du «Nord», une région rurale pauvre et très inégalitaire, à la tête du classement. On observe, encore ici, une corrélation négative entre les niveaux de pauvreté et d'inégalité. Il est alors important de considérer ces deux problèmes de façon simultanée afin de permettre un certain compromis entre les situations extrêmes (pauvreté versus inégalité).

5.2 Analyse statistique

Un des objectifs de ce mémoire était de mettre en relief l'impact du plan de sondage lors de l'estimation de la variance ou de l'erreur quadratique moyenne des estimateurs utilisés lors de l'analyse économique. Dans ce qui suit, nous utiliserons les résultats de l'étude afin de valider empiriquement la théorie statistique présentée au chapitre traitant de la méthodologie statistique. Nous démontrerons l'importance de prendre en compte le plan de sondage réel lors d'un sondage auprès des ménages d'une population. À notre connaissance, peu de chercheurs ont démontré l'impact du plan de sondage sur la précision des estimateurs (indicateurs économiques) à partir des valeurs de l'effet de plan. D'autre part, nous croyons que très peu d'auteurs ont présenté des résultats portant sur la validation des tests d'hypothèses utilisés dans le but de déterminer si l'écart entre deux courbes de Dominance stochastique (ou bien les valeurs des ordonnées de celles-ci) est statistiquement significatif.

5.2.1 La Pauvreté

Contrairement à ce qui est prédit par la théorie économique, nous remarquons que certains intervalles de confiance admettent des valeurs faiblement négatives pour les trois indices FGT lors de l'analyse des deux premiers groupes socio-économiques (les salariés). On doit alors fixer la borne inférieure à zéro. Ce problème peut, dans un premier temps, s'expliquer par le fait qu'en échantillons de faible taille, la loi normale est souvent une bien mauvaise approximation de la loi de probabilité des variables à l'étude. De plus, nous croyons que le problème peut être une conséquence de l'approximation linéaire de l'erreur échantillonnale de même que de l'hypothèse d'un plan supposé avec remise.²

L'analyse des effets de plan démontre très bien l'importance de prendre en compte le plan de sondage lors de l'estimation de la variance des estimateurs. Si on examine les résultats d'IMPS pour le niveau de vie per capita de la population du Burkina Faso, on retrouve une estimation de 72 395,04 f. CFA. L'estimation de l'erreur échantillonnale au niveau du plan réel (Plan 1 : stratification, deux degrés et probabilités de sélection inégales au degré 1) est de 1 838,57 f. CFA et l'effet de plan associé est de 4,33. Le fait de négliger totalement le plan de sondage lors du calcul de l'erreur conduira à une estimation de $\sqrt{(1838,57)^2 / 4,33} = 883,56$ (EAS). Se rapporter à la figure A.1 pour les différentes valeurs de l'effet de plan au niveau des dépenses totales et des dépenses moyennes réelles per capita. Comme la théorie échantillonnale a tendance à le prédire, le fait d'ignorer la structure du sondage, en faveur d'un plan aléatoire simple, conduit généralement à une sous-estimation de l'erreur échantillonnale.

5.2.2 Un domaine d'étude

Il est primordial de ne pas confondre les termes *strates* et *domaines d'étude*. Une strate peut être considérée comme un **sous-échantillon** indépendant. À la sous-section

²Nous rappelons qu'un plan avec remise permet d'obtenir des estimations de la variance qui sont conservatrices. On surestime alors la vraie variance.

«Stratification», nous avons présenté la variance d'un total comme étant la somme de la variance des totaux de chacune des strates. Dans le cas d'un domaine d'étude, ce résultat ne tient plus car il s'agit d'un **sous-ensemble** de l'échantillon et non d'un sous-échantillon. Comme exemple, prenons l'estimateur de la population totale. Pour le pays, l'estimation était de 9 392 567 individus et sa variance de $(360\ 141)^2 = 129\ 701\ 539\ 881$. Les variances estimées au niveau des strates sont (Plan 1) :

$$\begin{aligned}
 1 & : (303\ 643)^2 \\
 2 & : (87\ 277)^2 \\
 3 & : (103\ 137)^2 \\
 4 & : (114\ 977)^2 \\
 5 & : (36\ 326)^2 \\
 6 & : (47\ 811)^2 \\
 7 & : (49\ 217)^2.
 \end{aligned}$$

Nous avons alors :

$$\sum_{i=1}^7 (\quad)^2 = (360\ 141)^2.$$

Dans le cas des domaines d'étude tels les GSE on a :

$$\begin{aligned}
 1 & : (33\ 324)^2 \\
 2 & : (20\ 265)^2 \\
 3 & : (39\ 018)^2 \\
 4 & : (12\ 138)^2 \\
 5 & : (145\ 650)^2 \\
 6 & : (305\ 626)^2 \\
 7 & : (58\ 064)^2.
 \end{aligned}$$

et

$$\sum_{i=1}^7 (\quad)^2 = (348\ 114)^2 < (360\ 141)^2.$$

La différence fondamentale se situe au niveau de la structure du sondage. En fait, plusieurs unités peuvent appartenir à un même domaine d'étude sans toutefois appartenir à la même strate. De même, les unités d'une strate peuvent appartenir à différents domaines d'étude. La variance «globale» pour la population ne peut donc pas être décomposée dans le cas des domaines d'étude car ces derniers ne sont pas indépendants par rapport au plan de sondage.

5.2.3 Un profil de pauvreté

Il est maintenant possible de dresser un profil de pauvreté pour les deux zones géographiques à partir des tableaux A.1 et A.2. Le premier tableau donne une estimation de la population totale des zones rurale et urbaine alors que le second donne les estimations des indices FGT de ces zones. Soit la zone rurale (1) et la zone urbaine (2), alors :

$$\widehat{P}_\alpha = \frac{\widehat{N}_1}{\widehat{N}} \widehat{P}_\alpha^1 + \frac{\widehat{N}_2}{\widehat{N}} \widehat{P}_\alpha^2.$$

Sur le plan analytique, la contribution de la région rurale à l'indice de pauvreté global est alors de :

$$\frac{\widehat{N}_1}{\widehat{N}} \frac{\widehat{P}_\alpha^1}{\widehat{P}_\alpha} \%.$$

Dans le cas de P_0 on a la décomposition suivante :

$$\widehat{P}_0 = \frac{7\ 871\ 488}{9\ 392\ 567} \times 0,510 + \frac{1\ 521\ 079}{9\ 392\ 567} \times 0,104 = 0,444.$$

La contribution de la région rurale à la pauvreté totale est alors de :

$$\frac{\widehat{N}_1 \widehat{P}_0^1}{\widehat{N} \widehat{P}_0} = \frac{7\,871\,488}{9\,392\,567} \times \frac{0,510}{0,444} = 96,3 \%$$

La contribution de la région urbaine à la pauvreté totale est donc de 3,7%.

On peut faire de même pour les autres domaines d'étude (e.g. les strates d'analyse et les groupes socio-économiques).

5.2.4 L'inégalité

Comme dans le cas des indices de pauvreté (tableau A.3), on pourrait analyser les valeurs de l'effet de plan pour l'erreur des indices d'ATKINSON estimés à partir de l'échantillon. Les conclusions seraient similaires à celles que nous avons présentées lors de l'analyse de la pauvreté. Effectivement, nous avons observé des effets de plan inférieurs à l'unité au niveau de la strate d'analyse «Nord», des groupes «autres actifs» et «inactifs», ce qui laisse croire que l'effet stratification aurait pu renverser l'effet grappe. Les effets de plan pour les différents domaines d'étude varient entre 0,5 et 25, selon la valeur du paramètre d'aversion à l'inégalité. La plupart des \widehat{DF} obtenus dans cette étude (indices d'inégalité) prennent des valeurs supérieures à l'unité. En général, l'effet grappe dominera l'effet stratification et entraînera une réduction de la précision des estimateurs. On constate alors que le plan de sondage peut avoir un effet non négligeable sur la précision des estimations obtenues dans le cas de l'indice d'inégalité d'ATKINSON. Au tableau A.3, on retrouve quelques-uns des indices calculés dans le cadre du mémoire, de même que leur erreur échantillonnale.

5.2.5 La dominance stochastique

L'écart entre deux courbes de dominance stochastique. Dans un premier temps, nous avons choisi de tester l'écart entre les courbes d'incidence de la pauvreté des groupes

socio-économiques 6 et 7 («agriculteurs vivriers et inactifs») pour des seuils de pauvreté allant de 15 000 à 46 000 f. CFA. Suite à l'analyse des intervalles de confiance bilatéraux du tableau A.4, nous constatons que l'écart entre les courbes 6 et 7 n'est jamais significativement différent de zéro pour les seuils retenus. Pour les seuils en deçà de 15 000 f., il est peu probable que l'écart (toujours plus faible lorsque $z \rightarrow 0$) soit significatif. Il pourrait être intéressant de poursuivre les tests de dominance stochastique pour des seuils plus élevés quoique l'on ne pourra toutefois pas conclure qu'il y a Dominance du premier ordre car aucune courbe n'est initialement au dessus de l'autre pour certains z positifs. Dans ce cas-ci, l'alternance au niveau du signe (\pm) des estimations de l'écart démontre bien les croisements des deux courbes. Nous rappelons que la Condition de dominance du premier ordre stipule que la courbe d'incidence pour une situation doit être significativement supérieure à celle d'une autre situation pour $z \in [0, z_{\max}]$. On trouve des valeurs supérieures à l'unité pour l'effet de plan de chacune des estimations (de 2,97 à 7,54), comme le prédit généralement la théorie statistique. Se rapporter aux figures B.3 et B.4 ainsi qu'au tableau A.4.

Dans un deuxième temps, nous avons testé l'écart des courbes d'incidence au niveau des zones géographiques (1 pour la zone rurale et 2 pour la zone urbaine). Les intervalles de confiance bilatéraux montrent que les écarts sont significativement positifs pour les marges suivantes : $z \in [15\ 000, 23\ 571]$ f. CFA et $z \in [38\ 500, 46\ 000]$ f. CFA. Pour les seuils intermédiaires (23 571 à 38 500 f. CFA), l'écart entre la courbe de la zone rurale et celle de la zone urbaine n'est pas significativement différent de zéro. En posant l'hypothèse que les écarts sont toujours positifs pour des seuils allant de 0 à 15 000 f. CFA, on peut conclure qu'il y a Dominance du premier ordre pour $z_{\max} = 23\ 571$ f. CFA. Par contre, ce seuil maximal ne semble pas raisonnable comparativement aux seuils établis lors de l'Enquête Prioritaire (31 749 et 41 099 f. CFA). Nous n'avons pas testé l'écart pour les seuils inférieurs à 15 000 f., mais nous pensons que celui-ci deviendra non significatif bien avant que l'on atteigne la borne 0 f. CFA (les estimations de l'écart vont en décroissant au fur et à mesure que le seuil s'approche de 0 f. CFA). Encore une fois, les valeurs des effets de plan ne vont pas à l'encontre de la théorie statistique. Se rapporter aux figures B.5 et B.6 ainsi qu'au

tableau A.4.

L'ordonnée d'une courbe de dominance stochastique. La théorie statistique permet également de tester si l'ordonnée ($o_\alpha(z)$) d'une courbe de dominance stochastique est significativement différente d'une valeur spécifique (test bilatéral où $H_0 : o_\alpha(z) = c$ vs $H_1 : o_\alpha(z) \neq c$) pour un seuil de pauvreté donné. La valeur spécifiée doit par contre être strictement positive pour pouvoir effectuer le test bilatéral car il est évident, selon la théorie économique, que l'ordonnée d'une courbe de dominance stochastique est bornée inférieurement par zéro. Dans le cas du test unilatéral, on peut conclure que l'ordonnée est positive si l'estimation $\hat{o}_\alpha(z)$ est supérieure à la valeur critique $c^* = 1,645 \times \sqrt{v(\hat{o}_\alpha(z))}$ (1,645 est le quantile 0,95 de la loi normale centrée réduite). Les intervalles de confiance calculés pour la courbe d'incidence de la pauvreté de la zone urbaine (zone 2) ne permettent que de tester les hypothèses bilatérales ($H_0 : o_\alpha(z) = \tau$ vs $H_1 : o_\alpha(z) \neq \tau$ où $\tau > 0$). Il peut alors paraître étonnant de constater que les intervalles de confiance pour les seuils intermédiaires admettent des valeurs négatives tandis que pour les seuils extrêmes on obtient des intervalles qui appuient la théorie économique, indiquant que l'ordonnée d'une courbe de dominance ne peut prendre des valeurs négatives. Les intervalles qui admettent des valeurs négatives doivent être bornés inférieurement à zéro, ce qui peut sembler fausser le niveau de confiance, mais cela est une conséquence de l'application de la théorie asymptotique à des échantillons de taille finie. Nous avons aussi remarqué un saut dans la valeur de l'ordonnée au seuil 25 714 f. CFA, de même qu'une augmentation dramatique de son erreur échantillonnale. La plus forte hétérogénéité du niveau de vie des individus les plus nantis (dépenses réelles per capita supérieures à 25 714 f.), ceux-ci occupant surtout la zone urbaine et membres de ménages de plus petite taille, pourrait expliquer l'augmentation de l'erreur et également la forte augmentation de l'effet de plan. Se rapporter aux figures B.5 et B.6 de même qu'au tableau A.5.

Chapitre 6

Conclusion

En sciences économiques, les sondages auprès des ménages sont très souvent utilisés pour obtenir de l'information sur le niveau de vie des individus d'une population. Les économistes utiliseront cette information afin d'estimer des indices de pauvreté et d'inégalité qui permettront de dresser un profil de pauvreté et d'inégalité afin de repérer les régions ou les groupes les plus vulnérables de la population. Il sera ainsi possible de mieux concevoir des politiques de lutte contre la pauvreté et l'inégalité.

L'objectif premier de ce mémoire était de démontrer que le fait d'ignorer le plan de sondage lors d'une enquête statistique pouvait avoir un effet significatif sur l'estimation de la précision des estimateurs. Dans la réalité, quand on incorpore la stratification et l'effet grappe, les plans sondage sont beaucoup plus que des plans aléatoires simples. Pourtant, plusieurs chercheurs utilisent l'estimateur de variance naïf qui suppose que nous sommes en présence d'un échantillon aléatoire simple. L'estimateur obtenu sera souvent biaisé et l'estimation de sa variance ne sera pas valide. Le fait de négliger la stratification aura tendance à fournir une sous-estimation de la précision de l'estimateur tandis que le fait de négliger les degrés d'échantillonnage entraînera une surestimation de sa précision. Il est donc crucial de considérer le plan de sondage employé lors de l'estimation de la *valeur* et

de la variance d'un paramètre ou d'un indice. Souvent, on utilisera l'effet de plan («design effect») afin de mesurer l'impact du plan de sondage sur l'estimation de la variance.

La pertinence de nos recherches réside dans l'estimation de la variance de certains indices de pauvreté et d'inégalité couramment utilisés dans la littérature économique. Ces indices non linéaires ont dû être approximés par l'approche des séries de TAYLOR. Nous avons vu que l'estimation de la variance d'un ratio de variables aléatoires revenait à estimer une fonction de la variance et de la covariance de deux totaux, ce que nous estimons assez facilement à partir de la théorie échantillonnale.

Dans le cadre du programme MIMAP-formation dirigé par le CRÉFA et le CECI, nous avons estimé les principaux indices ainsi que leur variance à partir d'une base de données provenant d'un sondage mené auprès des ménages burkinabés au cours de la période allant d'octobre 1994 à janvier 1995. Ces indices peuvent être utilisés dans le but de dresser un profil de pauvreté et d'inégalité pour le Burkina Faso. Les résultats obtenus sont les suivants :

La pauvreté est surtout présente en milieu rural dans les strates «Centre nord» et «Centre sud». Cette région regroupe principalement des agriculteurs vivriers et rentiers et les ménages sont généralement de grande taille. L'inégalité caractérise davantage le milieu urbain. Les petites villes, ou la strate d'analyse «Autres villes», sont les plus touchées. En fait, on y retrouve différents types de travailleurs (artisans, commerçants, autres actifs) et beaucoup d'inactifs. De plus, les ménages y sont de plus petite taille. Les grandes villes (strate d'analyse Ouagadougou–Bobo-Dioulasso) sont les moins pauvres. Par contre, comme on y retrouve une forte masse de salariés publics bénéficiant d'une protection salariale, il n'est pas surprenant que l'inégalité y soit inférieure à celle des «Autres villes» et que le niveau de pauvreté y soit le plus faible au Burkina Faso. Nous croyons que les mesures de lutte contre la pauvreté devraient viser initialement la zone rurale («Centre nord» et «Centre sud») de même que certains groupes socio-économiques, tels les agriculteurs, les inactifs et les autres actifs. Ensuite, on pourra s'intéresser au problème de l'inégalité qui caractérise les secteurs plus fortunés.

Sur le plan des estimations, nous avons démontré qu'il était important de tenir compte de l'effet de plan lors de l'estimation de la variance, même si les poids échantillonnaires initiaux ont été établis en fonction de la structure réelle de l'enquête. Nous avons obtenu des effets de plan de l'ordre de 50. Par la suite, nous avons trouvé des estimations par intervalles allant à l'encontre de la théorie économique en admettant des valeurs négatives pour certains indices FGT. Ces derniers doivent être tronqués à zéro. Nous avons ensuite montré que les courbes de dominance stochastique présentent parfois des classements qui ne sont pas toujours statistiquement significatifs. Il est alors important de tester l'écart entre ces courbes sur l'intervalle des seuils de pauvreté admissibles.

En terminant, nous rappelons que l'analyse effectuée n'est pas exhaustive. Nous avons estimé les indices de pauvreté et d'inégalité les plus populaires pour certains domaines d'étude. Nous aimerions poursuivre l'étude en analysant des indicateurs d'inégalité plus complexes tels l'ordonnée de la courbe de LORENZ, le coefficient de GINI, les quantiles de la distribution du niveau de vie, etc. Ces indices s'estiment plus difficilement et l'estimation correcte de la variance, compte tenu du plan de sondage complexe, demande des techniques de linéarisation plus avancées. Peut-être aurons-nous recours aux techniques modernes de rééchantillonnage («Bootstrap, Jackknife, Balanced Repeated Replication», etc.) à des fins de comparaisons au niveau des estimations de la variance.

Chapitre 7

Bibliographie

- [1] ASSELIN L.-M. (1984) «TECHNIQUES DE SONDAGE avec applications à l'Afrique», Gaëtan Morin éd., 697p.
- [2] ATKINSON A.B. (1987) «On the Measurement of Poverty», Econometrica, 55,4, 749-764.
- [3] ATKINSON A.B. (1970) «On the Measurement of Inequality», Journal of Economic Theory, 2, 244-263.
- [4] BUREAU OF THE CENSUS, U.S. DEPARTMENT OF COMMERCE (1995) «CENVAR, Variance Calculation System et DATADICT (IMPS), User's Guides», janvier 1995, Washington, D.C., 163p.
- [5] CLARK S., HEMMING R. et ULPH D. (1981) «On Indices for Measurement Poverty», The Economic Journal, 91, 515-526.
- [6] COCHRAN W.G. (1977) «Sampling Techniques», Third Edition, John Wiley and Sons éd., 428p.

- [7] COWELL F.A. (1995) «Measuring Inequality», Second Edition, LSE Handbooks in Economics, 194p.
- [8] DAVIDSON R. et DUCLOS J.-Y. (2000) «Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality», Econometrica, 68 (6), Novembre, 1435-1465.
- [9] DEATON A. (1994) «The Analysis of Household Surveys», chapitres 1 et 3.1, Banque Mondiale, 99p.
- [10] FOSTER J.E. et SHORROCKS A.F. (1988) «Poverty Ordering», Econometrica, 56, 173-177.
- [11] FOSTER J.E. et al. (1984) «A Class of Decomposable Poverty Measures», Econometrica, 52, 761-776.
- [12] HOWES S. et LANJOUW J.O. (1998) «Does Sample Design Matter for Poverty Rate Comparisons?», Review of Income and Wealth, 44,1, 99-109.
- [13] HOWES S. et LANJOUW J.O. (1997) «Poverty Comparisons and Household Survey Design», LSMS Working Paper 129, Banque Mondiale, 35p.
- [14] INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE. (1996) «Rapport méthodologique de l'Enquête Prioritaire», 37p.
- [15] MINISTÈRE DE L'ÉCONOMIE, DES FINANCES ET DU PLAN ET INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE (1996) «Le Profil de Pauvreté au Burkina Faso», première édition, 170p.
- [16] MORIN H. (1993) «Théorie de l'échantillonnage», Les Presses de l'Université Laval, éd., 178p.

- [17] RAVALLION M. (1996) «Comparaisons de la Pauvreté : Concepts et Méthodes», LSMS 122, Banque Mondiale, 162p.
- [18] RAVALLION M. et BIDANI B. (1994) «How Robust is a Poverty Profile?», The World Bank Economic Review, 8, 75-102.
- [19] SEN A. (1976) «Poverty : An Ordinal Approach to Measurement», Econometrica, 44, 219-231.
- [20] SEN A. (1979) «Issues in the Measurement of Poverty», Scandinavian Journal of Economics, 81, 285-307.
- [21] WETTA C., KABORE T.S., BONZI K. B., SIKIROU S., SAWADOGO M., SOMDA P., (2000) «Le Profil d'inégalité et de pauvreté au Burkina Faso», CRÉFA, cahier de recherche 0002, 89p.

Annexe A

Tableaux

Tableau A.1
Le niveau de vie

Population totale du Burkina Faso					
Totaux	estimation	écart-type plan 1°	écart-type plan 2°	effet de plan	
Population	9 392 567	360 141	73 498,16	24,01	
Dépenses réelles	683 531 650 000	32 195 974 000	8 449 251 934	14,52	
Niveau de vie per capita	72 395,04	1 838,57	883,56	4,33	
Groupes socio-économiques					
GSE 6 (agriculteurs vivriers)					
Population	6 394 010,00	305 626	77 855,44	15,41	
Dépenses réelles	330 295 964 000	22 911 153 000	4 871 406 645	22,12	
Niveau de vie per capita	51 657,09	1 567,66	575,51	7,42	
GSE 7 (inactifs)					
Population	714 562	58 064	35 206,47	2,72	
Dépenses réelles	58 575 720 000	6 153 530 000		2,64	
Niveau de vie per capita	80 459,39	5 104,01	3 722,48	1,88	

Zones géographiques			
Zone 1 (zone rurale)			
Population	7 871 488	353 544	79 776,09
Dépenses réelles	437 423 391 000	28 677 350 000	5 388 804 055
Niveau de vie per capita	55 571	1 832,70	590,27
Zone 2 (zone urbaine)			
Population	1 521 079	68 616	42 553,84
Dépenses réelles	246 108 259 000	14 635 243 000	8 207 036 743
Niveau de vie per capita	159 460,40	6 722,00	3 830,21

Plan de sondage imposé au logiciel LIMPS

- Plan 1 : Stratification, deux degrés d'échantillonnage et probabilités de sélection inégales au premier degré
- Plan 2 : Échantillonnage aléatoire simple

Tableau A.2
Les indices FGT

Population totale du Burkina Faso						
Indice	estimation	écart-type plan 1*	écart-type plan 2*	effet de plan	nombre d'observations	
P0	0,444	0,016	0,008	4,33	8642	
P1	0,139	0,007	0,003	5,72	8642	
P2	0,060	0,003	0,001	5,05	8642	
Groupes socio-économiques						
GSE 6 (agriculteurs vivriers)						
P0	0,515	0,021	0,008	6,82	5 154	
P1	0,163	0,009	0,004	5,53	5 154	
P2	0,070	0,004	0,002	3,85	5 154	
GSE 7 (inactifs)						
P0	0,413	0,035	0,025	1,92	715	
P1	0,144	0,015	0,012	1,65	715	
P2	0,067	0,009	0,007	1,60	715	

Zones géographiques	
Zone 1 (zone rurale)	
P0	0,510 0,020 0,007 7,34 5 924
P1	0,161 0,008 0,003 5,94 5 924
P2	0,070 0,004 0,002 4,25 5 924
Zone 2 (zone urbaine)	
P0	0,104 0,013 0,009 1,88 2 718
P1	0,025 0,004 0,003 1,67 2 718
P2	0,009 0,002 0,002 1,32 2 718

Plan de sondage imposé au logiciel JMPS

- * Plan 1 : Stratification, deux degrés d'échantillonnage et probabilités de sélection inégales au premier degré
- * Plan 2 : Échantillonnage aléatoire simple

Tableau A.3
L'indice d'Atkinson

Population totale du Burkina Faso						
epsilon	estimation	écart-type plan 1*	écart-type plan 2*	effet de plan	nombre d'observations	
0,25	0,0996	0,0046	0,002552	3,25	8642	
0,50	0,1801	0,0074	0,0040	3,48	8642	
1,05	0,3112	0,0102	0,005376	3,60	8642	
2	0,4549	0,0110	0,006372	2,98	8642	
Groupes socio-économiques						
GSE 6 (agriculteurs vivriers)						
0,25	0,0516	0,0028	0,00085	9,35	5 154	
0,50	0,0982	0,0045	0,001513	8,85	5 154	
1,05	0,1871	0,0068	0,002694	6,37	5 154	
2	0,3104	0,0081	0,004884	2,75	5 154	
GSE 7 (inactifs)						
0,25	0,1031	0,0010	0,0010	0,94	715	
0,50	0,1892	0,0015	0,001508	0,99	715	
1,05	0,3372	0,0019	0,001828	1,08	715	
2	0,5033	0,0020	0,001768	1,28	715	

Zones géographiques					
Zone 1 (zone rurale)					
0,25	0,0670	0,0030	0,0010	8,27	5 924
0,50	0,1239	0,0052	0,001883	7,63	5 924
1,05	0,2244	0,0077	0,0032	5,79	5 924
2	0,3520	0,0091	0,005236	3,02	5 924
Zone 2 (zone urbaine)					
0,25	0,0913	0,0035	0,002298	2,32	2 718
0,50	0,1699	0,0056	0,003528	2,52	2 718
1,05	0,3104	0,0076	0,004518	2,83	2 718
2	0,4833	0,0073	0,004173	3,06	2 718

Plan de sondage imposé au logiciel IMPS

* Plan 1 : Stratification, deux degrés d'échantillonnage et probabilités de sélection inégales au premier degré

* Plan 2 : Échantillonnage aléatoire simple

Tester l'écart entre deux courbes de dominance stochastique

Tableau A.4

Groupes socio-économiques 6 et 7						
	Seuil de pauvreté	estimation de l'écart	écart-type plan 1*	I. C. (95%)		effet de plan
				b. inf.	b. sup.	
GSE 6 - GSE 7 (agriculteurs vivriers - Inactifs)						
1	15 000	0,003	0,008	-0,012	0,019	3,72
2	17 143	-0,040	0,039	-0,117	0,037	7,54
3	19 286	-0,047	0,044	-0,133	0,040	6,92
4	21 429	-0,044	0,044	-0,130	0,042	5,71
5	23 571	-0,031	0,046	-0,120	0,059	5,17
6	25 714	-0,001	0,046	-0,091	0,089	5,17
7	27 857	0,032	0,043	-0,052	0,116	3,66
8	30 000	0,032	0,051	-0,068	0,133	3,85
9	32 071	0,057	0,050	-0,041	0,154	3,52
10	34 214	0,036	0,050	-0,063	0,134	2,97
11	36 357	-0,005	0,066	-0,134	0,123	4,25
12	38 500	-0,023	0,067	-0,108	0,154	4,38
13	40 643	0,027	0,069	-0,109	0,162	4,68
14	42 786	0,046	0,069	-0,088	0,180	4,58
15	44 929	0,031	0,071	-0,108	0,170	4,87
16	46 000	0,042	0,070	-0,096	0,180	4,81

Zones géographiques 1 et 2						
Zone 1 - Zone 2 (zone rurale - zone urbaine)						
1	15 000	0,009	0,003	0,003	0,015	1,06
2	17 143	0,029	0,008	0,014	0,044	4,27
3	19 286	0,045	0,010	0,026	0,064	2,59
4	21 429	0,067	0,012	0,043	0,091	2,12
5	23 571	0,103	0,016	0,073	0,134	2,84
6	25 714	-0,005	0,112	-0,224	0,215	3,84
7	27 857	0,042	0,112	-0,178	0,262	3,87
8	30 000	0,091	0,112	-0,129	0,311	3,91
9	32 071	0,125	0,113	-0,096	0,345	3,96
10	34 214	0,163	0,112	-0,057	0,383	3,97
11	36 357	0,196	0,112	-0,023	0,415	4,00
12	38 500	0,219	0,109	0,006	0,432	3,86
13	40 643	0,248	0,108	0,036	0,461	3,89
14	42 786	0,266	0,107	0,055	0,476	3,86
15	44 929	0,288	0,107	0,080	0,497	3,88
16	46 000	0,295	0,106	0,087	0,503	3,88

Note: Les flèches (<) signifient que l'écart est significativement positif.

Tableau A.5

Tester l'ordonnée d'une courbe de dominance stochastique

	Seuil de pauvreté	estimation de l'ordonnée	écart-type plan 1	I. C. (95%) b. inf.	b. sup.	effet de plan
Zone géographique 2 (zone urbaine)						
1	15 000	0,002	0,001	0,000	0,004	0,14
2	17 143	0,002	0,001	0,000	0,005	0,15
3	19 286	0,006	0,003	0,000	0,012	0,29
4	21 429	0,013	0,004	0,005	0,020	0,29
5	23 571	0,016	0,004	0,007	0,024	0,30
6	25 714	0,150	0,111	-0,067	0,368	3,72
7	27 857	0,157	0,110	-0,059	0,372	3,69
8	30 000	0,164	0,109	-0,050	0,377	3,67
9	32 071	0,169	0,108	-0,043	0,381	3,65
10	34 214	0,177	0,107	-0,033	0,388	3,61
11	36 357	0,190	0,106	-0,016	0,397	3,55
12	38 500	0,207	0,103	0,004	0,410	3,49
13	40 643	0,216	0,102	0,016	0,417	3,46
14	42 786	0,230	0,101	0,033	0,427	3,39
15	44 929	0,243	0,099	0,049	0,437	3,34
16	46 000	0,249	0,098	0,057	0,442	3,31

Note: Tronquer la borne inférieure de l'intervalle de confiance à zéro aux endroits marqués (?).

Annexe B

Figures

Figure B.1

Courbes de l'incidence de la pauvreté pour les huit strates d'analyse (20000 - 80000 f. CFA)

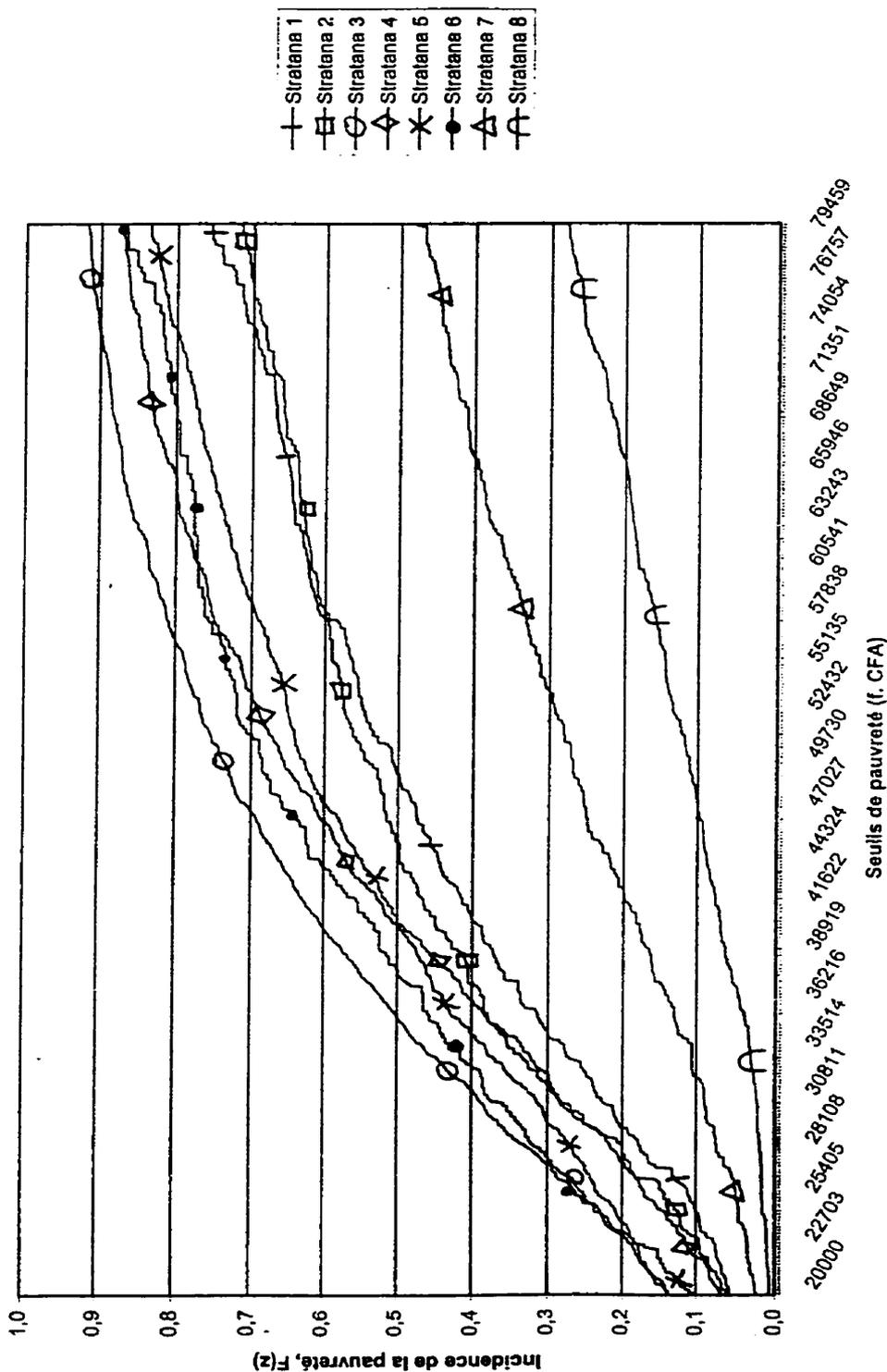


Figure B.2

Courbes de l'incidence de la pauvreté pour les huit strates d'analyse (15000 - 30000 f. CFA)

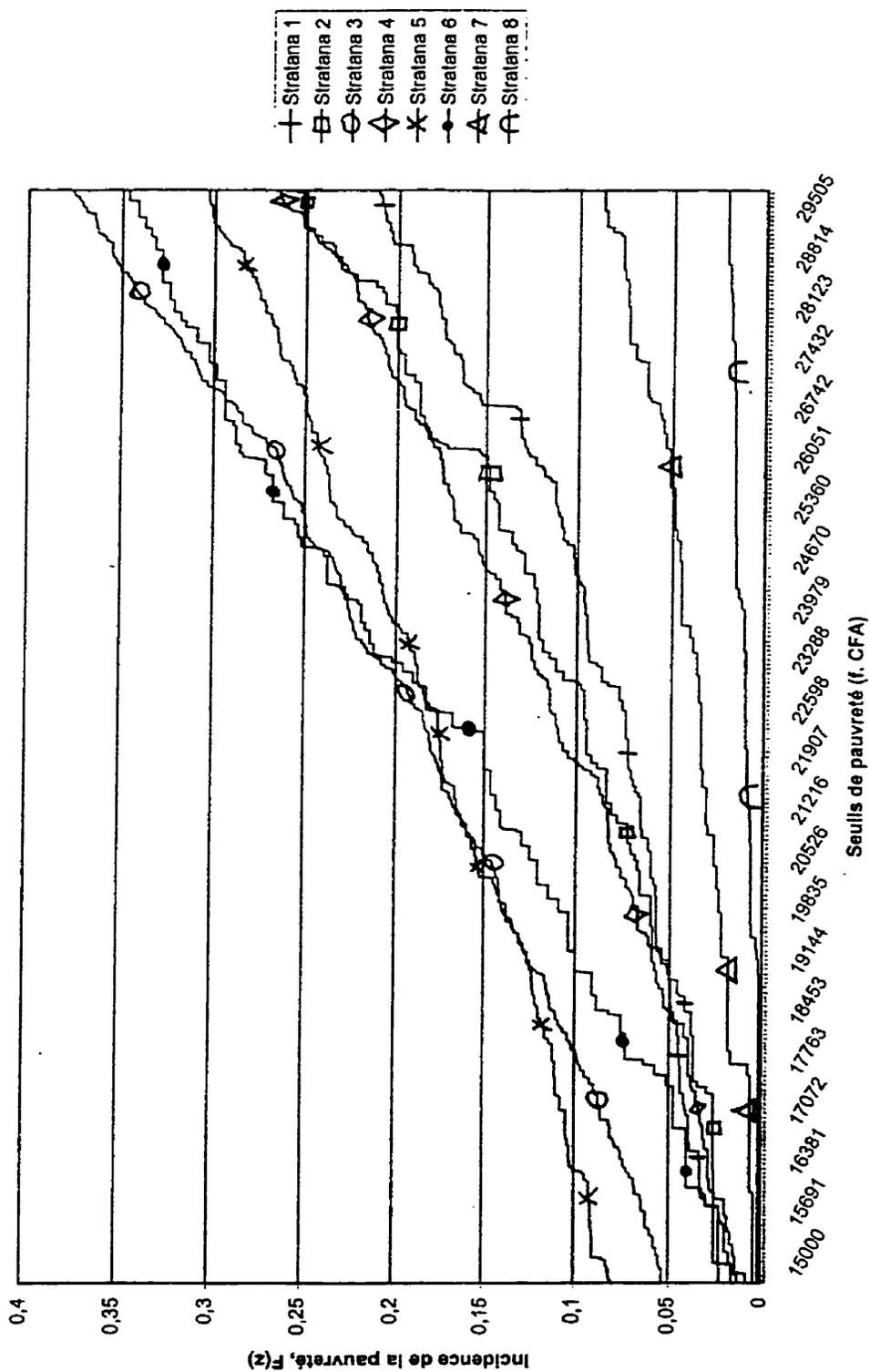


Figure B.3

Courbes de l'incidence de la pauvreté pour les sept groupes socio-économiques
(20000 - 80000 f. CFA)

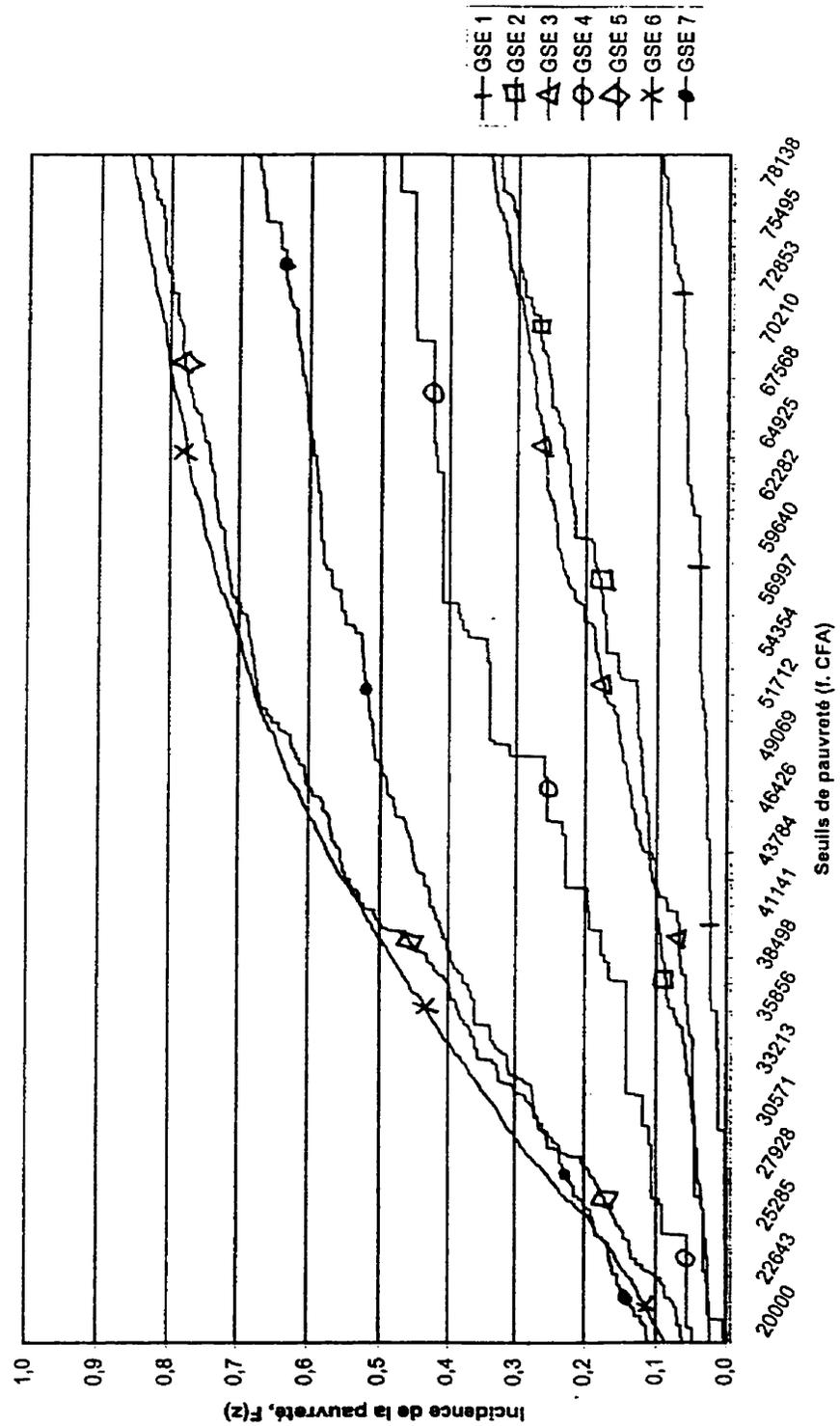


Figure B.4

Courbes de l'incidence de la pauvreté pour les sept groupes socio-économiques
(15000 - 30000 f. CFA)

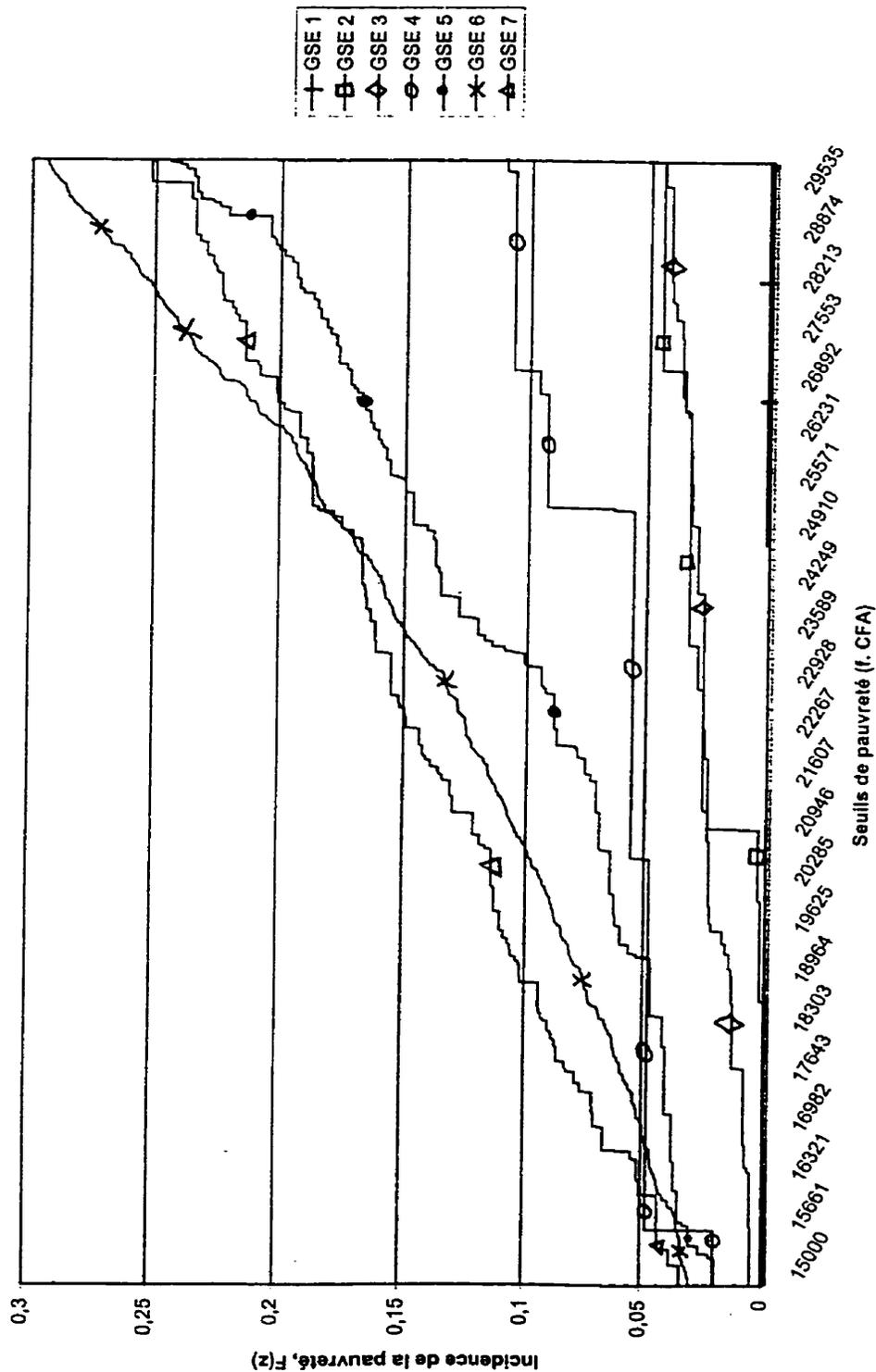


Figure B.5
 Courbes de l'incidence de la pauvreté pour les deux zones géographiques
 (20000 - 80000 f. CFA)

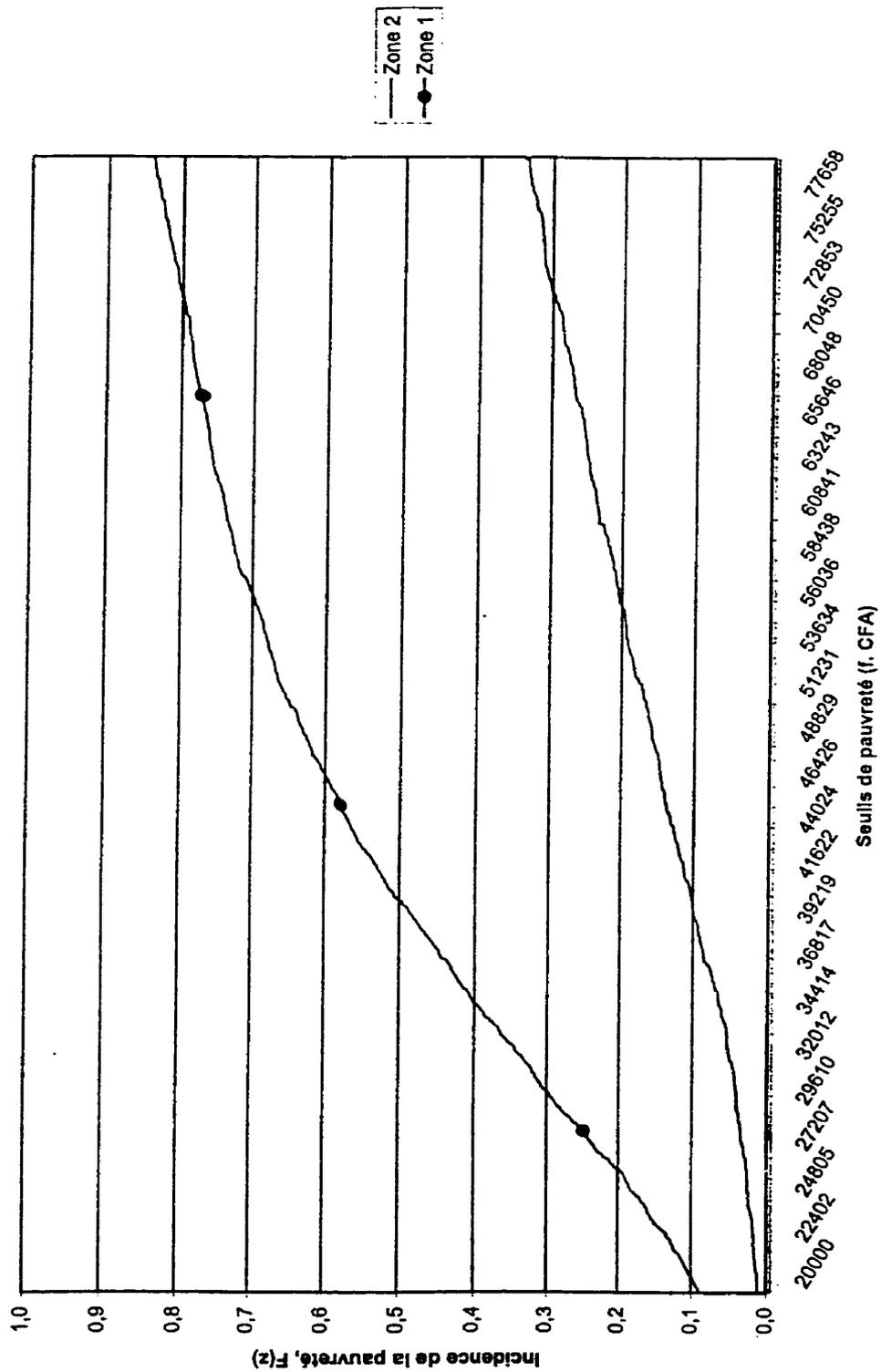


Figure B.6

Courbes de l'incidence de la pauvreté pour les deux zones géographiques
(15000 - 30000 f. CFA)

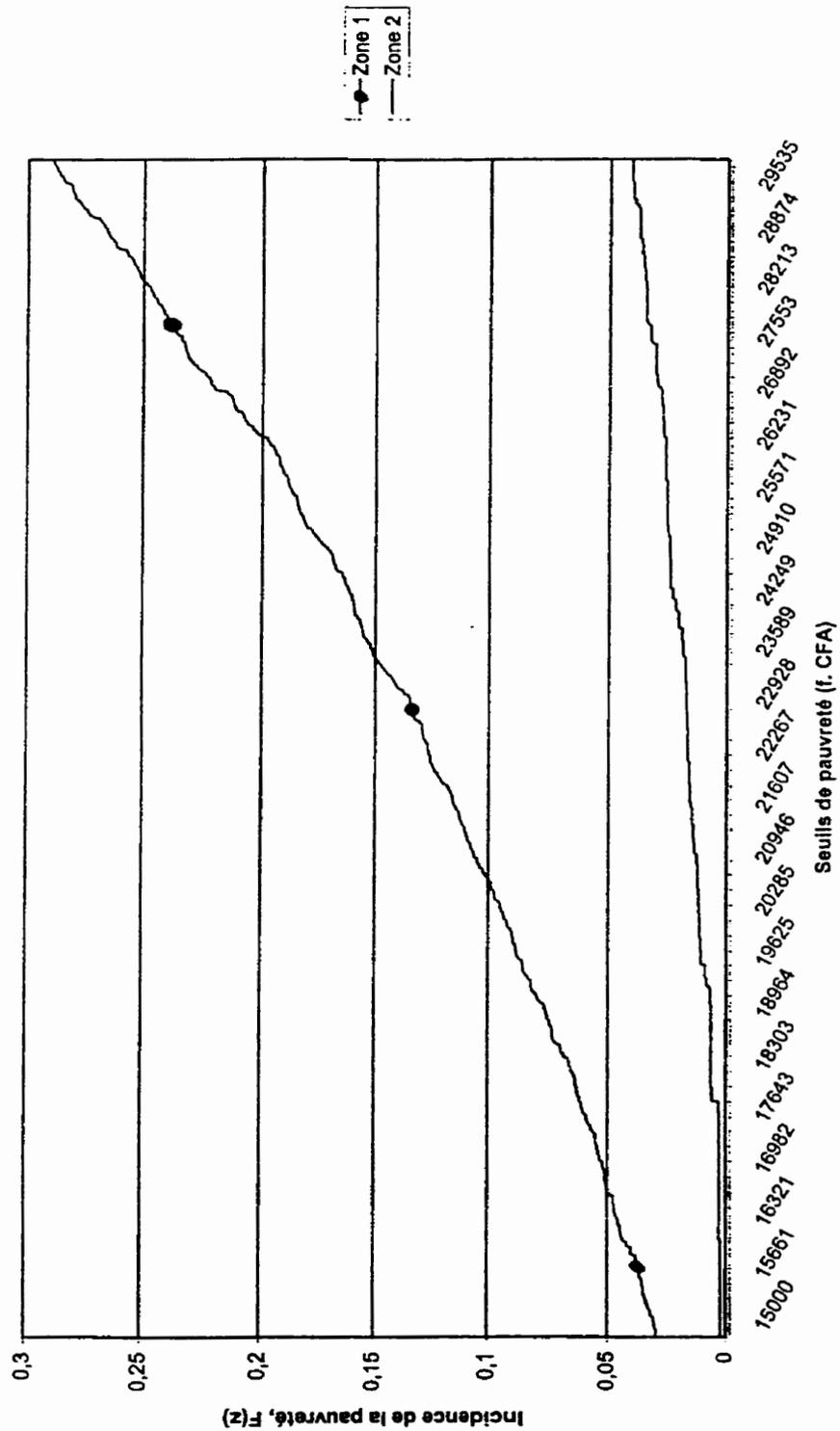


Figure B.7

Courbes du déficit de la pauvreté pour les groupes socio-économiques 6 et 7
(20000 - 80000 f. CFA)

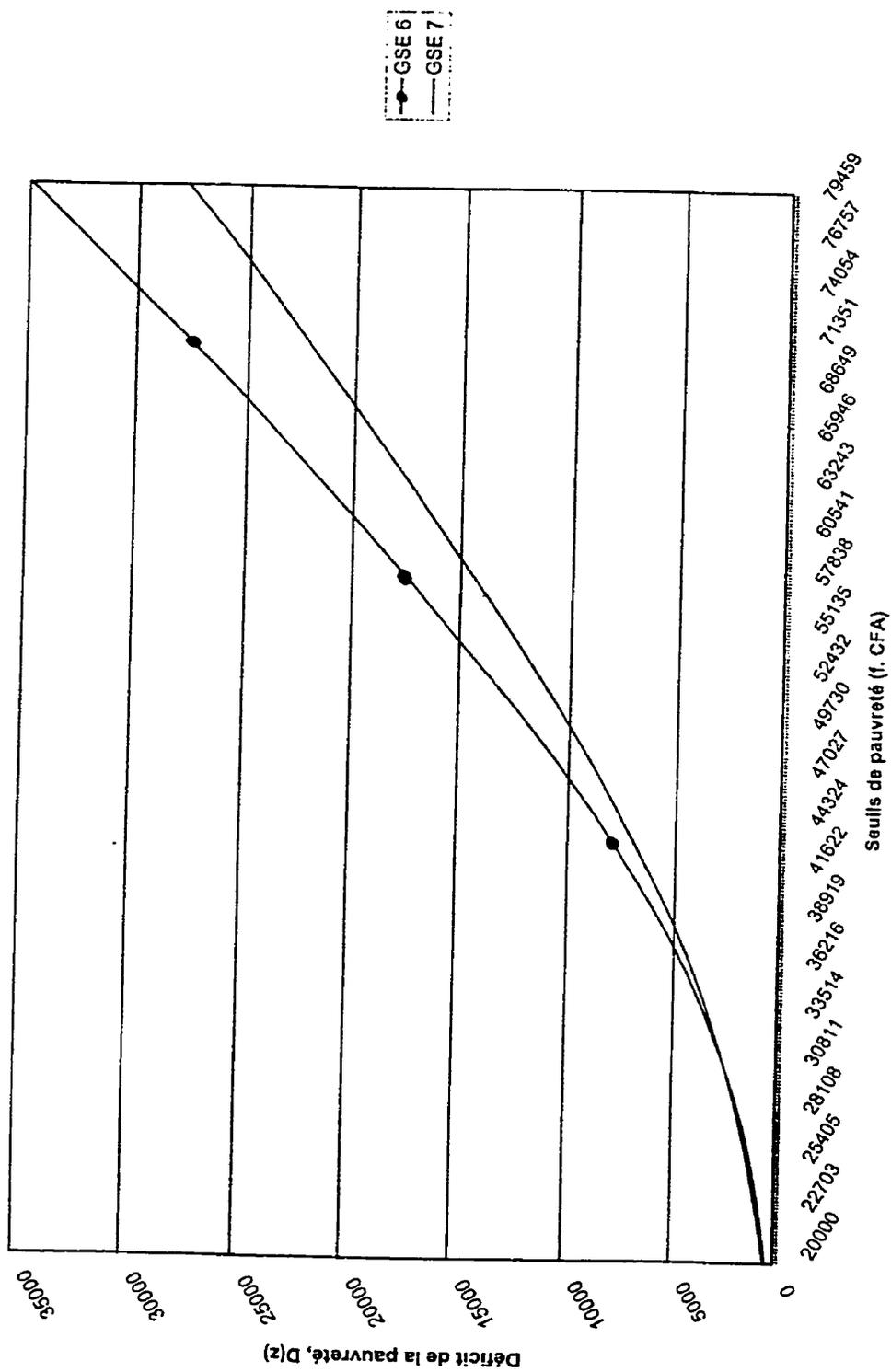


Figure B.8

Courbes du déficit de la pauvreté pour les groupes socio-économiques 6 et 7
(15000 - 30000 f. CFA)

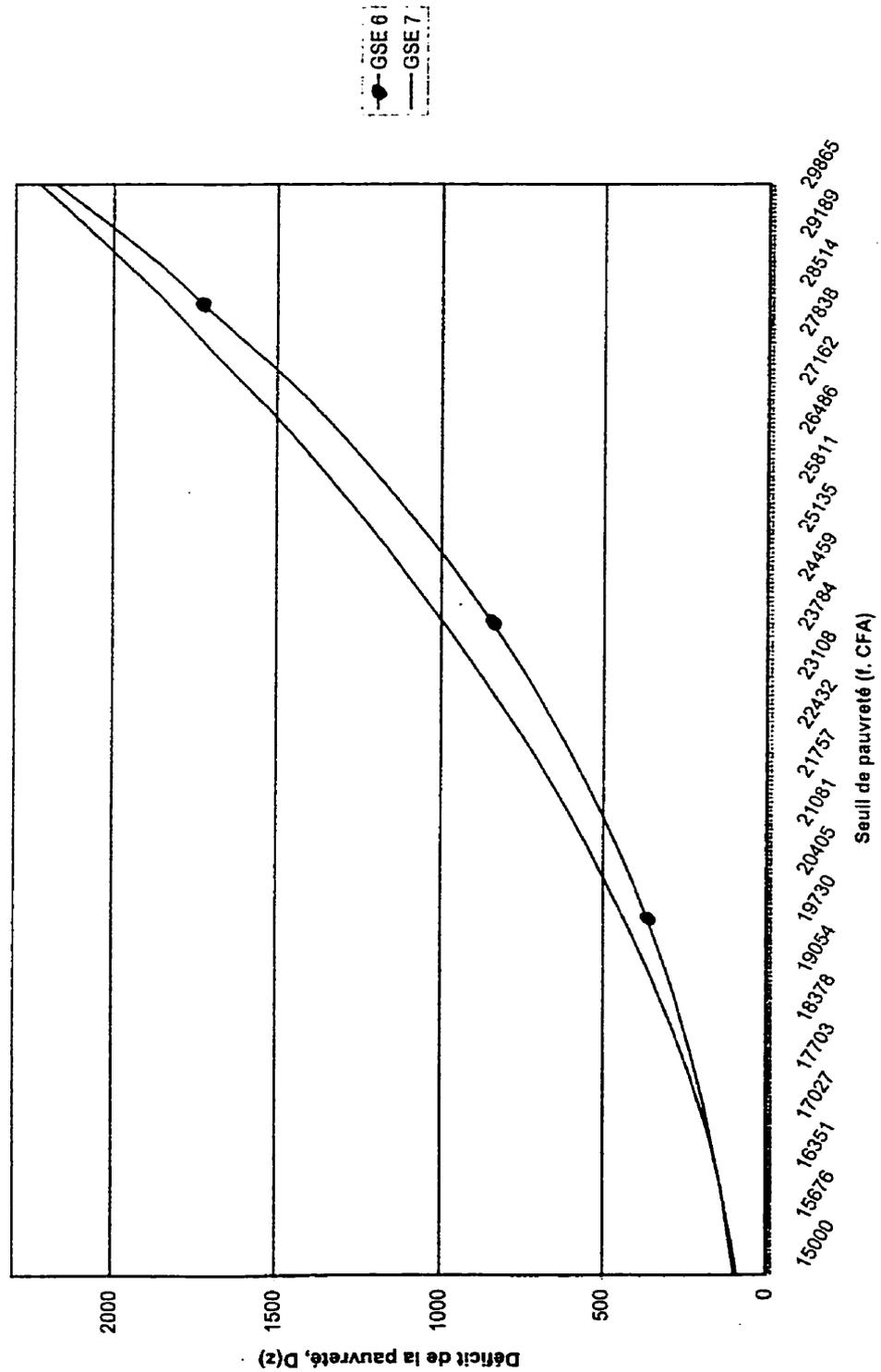


Figure B.9

Courbes de l'intensité de la pauvreté pour les groupes socio-économiques 6 et 7 (20000 - 80000 f. CFA)

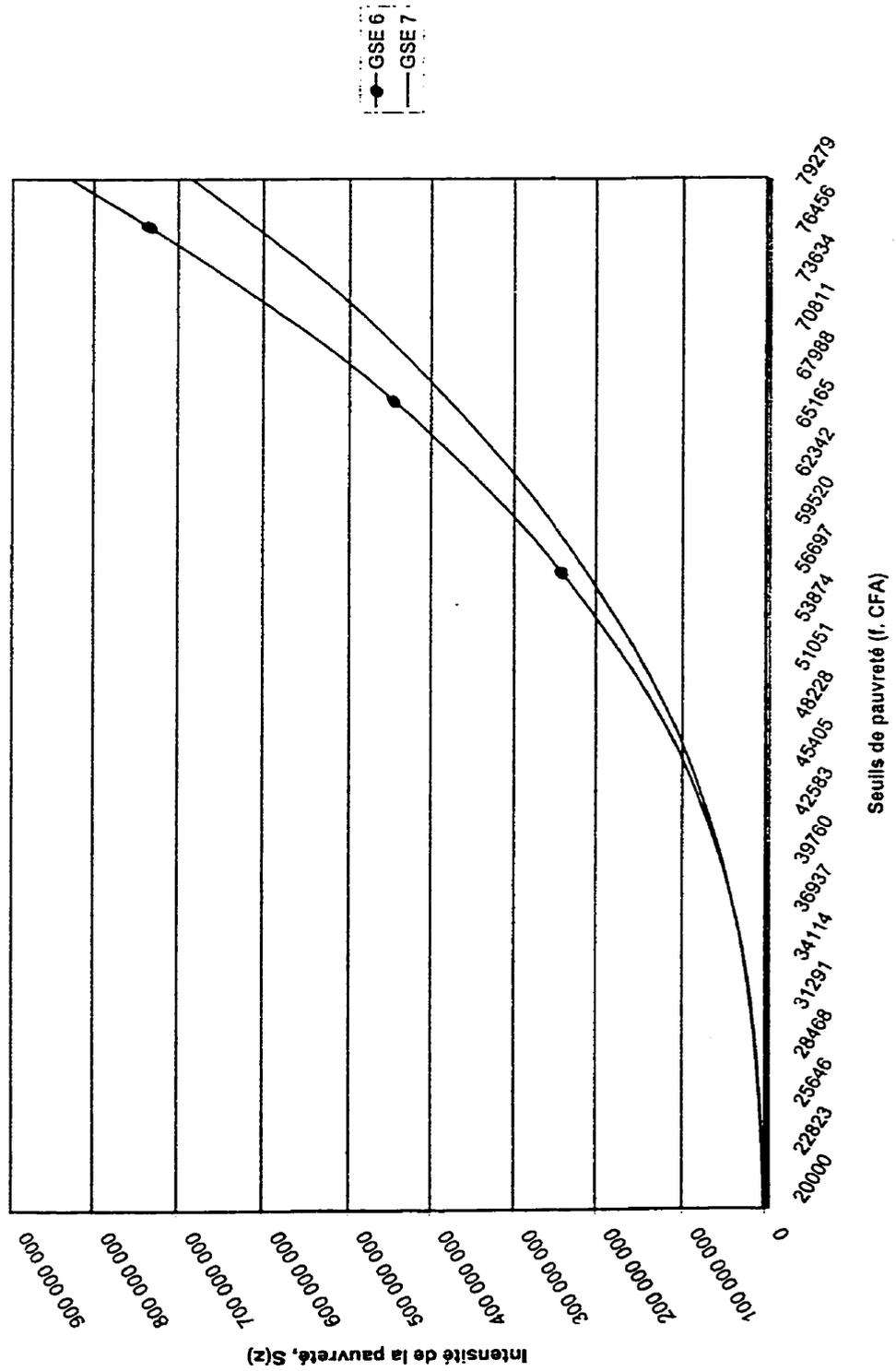
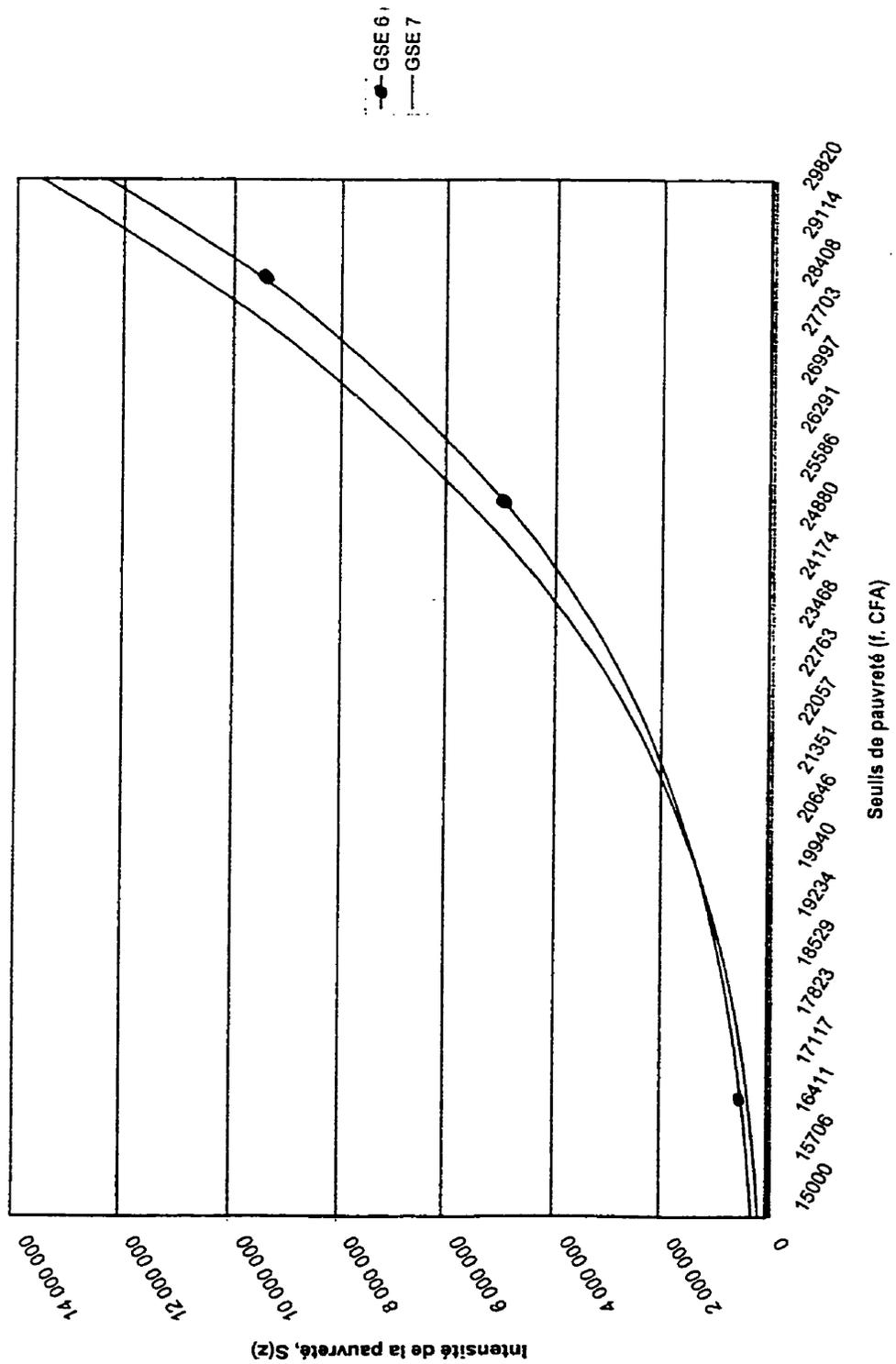


Figure B.10

Courbes de l'intensité de la pauvreté pour les groupes socio-économiques 6 et 7 (15000 - 30000 f. CFA)



Annexe C

Compléments théoriques

C.1 La linéarisation des estimateurs non linéaires

C.1.1 L'indice FGT

L'indice FGT s'estime comme un quotient de deux totaux. Le numérateur est une estimation de la somme des mesures de pauvreté individuelles pour l'ensemble de la population, alors que le dénominateur est une estimation de la population totale (nombre d'individus). Soit un quotient de variables aléatoires, sous forme de totaux :

$$R = \frac{Y}{X}, \quad (\text{C.88})$$

$$\widehat{R} = \frac{\widehat{Y}}{\widehat{X}}. \quad (\text{C.89})$$

Prenons la fonction logarithmique de \widehat{R} :

$$\ln \widehat{R} \simeq \ln R + \left. \frac{\partial \ln r}{\partial r} \right|_{r=R} (\widehat{R} - R). \quad (\text{C.90})$$

Il s'agit tout simplement de l'expansion de TAYLOR dans le cas d'une fonction simple :

$$\ln \widehat{R} - \ln R \simeq \frac{\widehat{R} - R}{R}, \quad (\text{C.91})$$

où $\ln R$ est un terme non aléatoire. Nous avons que :

$$\begin{aligned}
E(\ln \widehat{R} - \ln R)^2 &\simeq \frac{1}{\widehat{R}^2} E(\widehat{R} - R)^2 \\
&\Downarrow \\
Var(\ln \widehat{R}) &\simeq \frac{1}{\widehat{R}^2} Var(\widehat{R}) \\
&\Downarrow \\
Var(\widehat{R}) &\simeq R^2 \cdot Var(\ln \widehat{R}) \\
&= \frac{Y^2}{X^2} \begin{bmatrix} Var(\ln \widehat{Y}) + Var(\ln \widehat{X}) \\ -2 \cdot COV(\ln \widehat{X}, \ln \widehat{Y}) \end{bmatrix}. \tag{C.92}
\end{aligned}$$

Or,

$$\ln(\widehat{Y}) \simeq \ln Y + \frac{(\widehat{Y} - Y)}{Y}, \tag{C.93}$$

$$Var(\ln \widehat{Y}) \simeq \frac{1}{Y^2} Var(\widehat{Y}) \tag{C.94}$$

et

$$COV(\ln \widehat{X}, \ln \widehat{Y}) \simeq \frac{1}{X \cdot Y} COV(\widehat{X}, \widehat{Y}). \tag{C.95}$$

On obtient alors le résultat :

$$\begin{aligned}
Var(\widehat{R}) &\simeq \frac{Y^2}{X^2} \left[\frac{Var(\widehat{Y})}{\widehat{Y}^2} + \frac{Var(\widehat{X})}{\widehat{X}^2} - \frac{2 \cdot COV(\widehat{X}, \widehat{Y})}{\widehat{X} \cdot \widehat{Y}} \right] \\
&= \frac{Y^2}{X^2} \cdot \frac{1}{Y^2} \left[\frac{Y^2 \cdot Var(\widehat{Y})}{Y} + \frac{Y^2 \cdot Var(\widehat{X})}{X} - \frac{2 \cdot Y^2 \cdot COV(\widehat{X}, \widehat{Y})}{X \cdot Y} \right] \\
&= \frac{1}{X^2} \left[Var(\widehat{Y}) + \widehat{R}^2 \cdot Var(\widehat{X}) - 2 \cdot \widehat{R} \cdot COV(\widehat{X}, \widehat{Y}) \right]. \tag{C.96}
\end{aligned}$$

C'est le résultat que nous avons présenté à la sous-section 3.1.2.

C.1.2 L'indice d'Atkinson

Cet indice s'exprime à la fois comme un produit et un quotient d'estimateurs. Nous l'avons exprimé ainsi :

$$A_\varepsilon = 1 - \left[\frac{Z^a}{Y \cdot X^{a-1}} \right], \quad (\text{C.97})$$

$$\widehat{A}_\varepsilon = 1 - \left[\frac{\widehat{Z}^a}{\widehat{Y} \cdot \widehat{X}^{a-1}} \right]. \quad (\text{C.98})$$

Il faut donc linéariser l'expression entre crochets en posant :

$$\widehat{\theta} = \frac{\widehat{Z}^a}{\widehat{Y} \cdot \widehat{X}^{a-1}}, \quad (\text{C.99})$$

$$\ln \widehat{\theta} \simeq \ln \theta + \left. \frac{\partial \ln r}{\partial r} \right|_{r=\theta} (\widehat{\theta} - \theta), \quad (\text{C.100})$$

$$\text{Var}(\widehat{\theta}) \simeq \theta^2 \text{Var}(\ln \widehat{\theta}). \quad (\text{C.101})$$

Dans ce cas-ci, on écrit :

$$\begin{aligned} \ln \widehat{\theta} &= a \ln \widehat{Z} - (a-1) \ln \widehat{X} - \ln \widehat{Y} \\ &\Downarrow \\ \ln \widehat{\theta} &\simeq \ln \theta + a \left. \frac{\partial \ln e}{\partial e} \right|_{e=Z} (\widehat{Z} - Z) - (a-1) \left. \frac{\partial \ln p}{\partial p} \right|_{p=X} (\widehat{X} - X) \\ &\quad - \left. \frac{\partial \ln t}{\partial t} \right|_{t=Y} (\widehat{Y} - Y) \\ &\Downarrow \\ \ln \widehat{\theta} - \ln \theta &\simeq \frac{a}{Z} (\widehat{Z} - Z) - \frac{(a-1)}{X} (\widehat{X} - X) - \frac{(\widehat{Y}-Y)}{Y} \\ &= a \frac{\widehat{Z}}{Z} - (a-1) \frac{\widehat{X}}{X} - \frac{\widehat{Y}}{Y}. \end{aligned} \quad (\text{C.102})$$

Prenons la variance de cette dernière expression avec $\ln \theta$ constant :

$$\text{Var}(\ln \widehat{\theta}) = \text{Var}(\ln \widehat{\theta} - \ln \theta). \quad (\text{C.103})$$

Or, on trouve que

$$\text{Var}(\ln \hat{\theta}) \simeq \text{Var} \left(a \frac{\hat{Z}}{\hat{Z}} - (a-1) \frac{\hat{X}}{\hat{X}} - \frac{\hat{Y}}{\hat{Y}} \right), \quad (\text{C.104})$$

où $\left(a \frac{\hat{Z}}{\hat{Z}} - (a-1) \frac{\hat{X}}{\hat{X}} - \frac{\hat{Y}}{\hat{Y}} \right) = \sum_s k_s$ est une somme de variables aléatoires tel que démontré à la sous-section 3.2.4. Finalement,

$$\text{Var}(\hat{\theta}) \simeq \theta^2 \text{Var}(\ln \hat{\theta}). \quad (\text{C.105})$$

Il s'agit alors d'estimer la variance de la somme des k_s pour obtenir une estimation de la variance du logarithme de $\hat{\theta}$ en prenant le soin de remplacer les dénominateurs (X, Y, Z) par leur valeur estimée à partir de l'échantillon.

C.2 Estimation de la variance d'un estimateur de total

Soit un plan de sondage à deux degrés et probabilités de sélection inégales au premier degré (indifféremment des probabilités de sélection au deuxième degré). En supposant que l'échantillonnage s'est fait avec remise au degré 1, on peut facilement obtenir un estimateur de la variance du total.

Un certain nombre d'UPE (d) seront sélectionnées avec remise. Pour chacune d'elles, nous avons une valeur de l'estimateur d'HORVITZ-THOMPSON pour le total de la population ($\frac{\hat{Y}_c}{\pi_c} = \frac{\hat{Y}_c}{d p_c} = \text{tot}_c$). L'estimateur HT du total est simplement

$$\hat{Y}_{HT} = \frac{1}{d} \sum_{c=1}^d \text{tot}_c \quad (\text{C.106})$$

et l'estimateur de la variance (avec remise au niveau des UPE) de \widehat{Y}_{HT} sera simplement

$$v(\widehat{Y}_{HT}) = \frac{s_{tot}^2}{d} = \frac{1}{d} \left[\frac{\sum_{c=1}^d (tot_c - \widehat{Y}_{HT})^2}{d-1} \right]. \quad (C.107)$$

C.3 Comparaison de l'échantillonnage stratifié proportionnel à l'échantillonnage simple

Pour l'échantillonnage simple avec remise,¹ $Var(\bar{y}) = \frac{S_y^2}{m}$. Dans le cas de l'échantillonnage stratifié proportionnel ($m_h = mW_h$), nous avons $Var(\bar{y}_{stp}) = \frac{1}{m} \sum_{h=1}^L W_h S_{yh}^2$. Nous pouvons dire que l'échantillonnage stratifié proportionnel sera préféré à l'échantillonnage aléatoire simple si $Var(\bar{y}_{stp}) \leq Var(\bar{y})$, ou si $\sum_{h=1}^L W_h S_{yh}^2 \leq S_y^2$. Avec plusieurs manipulations, il est possible d'exprimer S_y^2 comme suit :²

$$S_y^2 \simeq \sum_{h=1}^L W_h S_{yh}^2 + \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2. \quad (C.108)$$

Comme le dernier terme est toujours positif, $\sum_{h=1}^L W_h S_{yh}^2 \leq S_y^2$ et $Var(\bar{y}_{stp}) \leq Var(\bar{y})$. Toutefois, il est possible d'obtenir des estimations de la variance plus élevées dans le cas de l'échantillonnage stratifié proportionnel car :

$$s_y^2 = \sum_{h=1}^L W_h s_{yh}^2 + \frac{1}{(m-1)} \left[\sum_{h=1}^L m_h (\bar{y}_h - \bar{y})^2 - \sum_{h=1}^L (1 - W_h) s_{yh}^2 \right]. \quad (C.109)$$

¹Lors de l'Enquête Prioritaire, les UPE ont été sélectionnées avec remise dans chacune des strates. Très souvent, les statisticiens utilisent un plan stratifié **proportionnel** où $m_h = mW_h$.

²Se référer à Morin (1993) pour les détails de la preuve.

L'expression entre crochets peut alors être négative, surtout si les strates sont hétérogènes et si les m_h sont petits. Parfois, la stratification peut mener à une surestimation de la variance.