Organisation internationale du Travail

L'échantillonnage pour les enquêtes auprès des ménages sur le travail des enfants

Programme d'information statistique et de suivi sur le travail des enfants (SIMPOC)

Programme international pour l'abolition du travail des enfants (IPEC)

Vijay VERMA

L'échantillonnage pour les enquêtes auprès des ménages sur le travail des enfants

Vijay VERMA

Programme
international
pour l'abolition
du travail des enfants
(IPEC)

Organisation international du Travail (OIT)

Copyright © Organisation internationale du Travail 2009
Première édition 2009

Les publications du Bureau international du Travail jouissent de la protection du droit d'auteur en vertu du protocole n° 2, annexe à la Convention universelle pour la protection du droit d'auteur. Toutefois, de courts passages pourront être reproduits sans autorisation, à la condition que leur source soit dûment mentionnée. Toute demande d'autorisation de reproduction ou de traduction devra être envoyée à l'adresse suivante: Publications du BIT (Droits et licences), Bureau international du Travail, CH-1211 Genève 22, Suisse, ou par courriel: pubdroit@ilo.org. Ces demandes seront toujours les bienvenues.

Bibliothèques, institutions et autres utilisateurs enregistrés auprès d'un organisme de gestion des droits de reproduction ne peuvent faire des copies qu'en accord avec les conditions et droits qui leurs ont été octroyés. Visitez le site www.ifrro.org afin de trouver l'organisme responsable de la gestion des droits de reproduction dans votre pays.

IPEC, V. Verma

L'échantillonnage pour les enquêtes auprès des ménages sur le travail des enfants / IPEC, Vijay Verma; Organisation internationale du Travail, Programme international pour l'abolition du travail des enfants - Genève: OIT, 2009 - 284 p.

ISBN: 978-2-2-222532-3 (Print), 978-92-2-221504-1(Web PDF); 978-92-2-221505-8 (CD-Rom)

International Labour Office; ILO International Programme on the Elimination of Child Labour

guide / travail des enfants / enquête / échantillon / méthode statistique / pays en développement - 13.01.2

Egalement disponible en anglais: *Sampling for household-based surveys of child labour*, ISBN 978-92-2-121503-5 (Print), 978-92-2-121504-2 (Web PDF), 978-92-2-121505-9 (CD-Rom), Genève, 2008., et en espagnol: *Muestreo para las encuestas de hogares sobre trabajo infantil*, ISBN 978-92-2-322533-9 (Print), 978-92-2-321504-0 (Web PDF); 978-92-2-321505-7 (CD-Rom), Genève, 2009.

Données de catalogage du BIT

Note

Cette publication a été élaborée par Vijay Verma pour l'IPEC et coordonnée par Mustafa Hakki Ozel du Bureau de l'IPEC à Genève.

Cette publication de l'OIT a été financée par le ministère du Travail des Etats-Unis (*Department of labour)* (Projet INT/03/61/USA).

Les opinions qui y sont exprimées ne reflètent pas nécessairement les points de vue ou les politiques de ce ministère. Le fait que des marques commerciales, des produits commerciaux ou des organismes y soient mentionnés ne signifie pas non plus qu'ils sont cautionnés par le Gouvernement des Etats-Unis.

Les désignations utilisées dans les publications du BIT, qui sont conformes à la pratique des Nations Unies, et la présentation des données qui y figurent n'impliquent de la part du Bureau international du Travail aucune prise de position quant au statut juridique de tel ou tel pays, zone ou territoire, ou de ses autorités, ni quant au tracé de ses frontières.

Les articles, études et autres textes signés n'engagent que leurs auteurs et leur publication ne signifie pas que le Bureau international du Travail souscrit aux opinions qui y sont exprimées.

La mention ou la non-mention de telle ou telle entreprise ou de tel ou tel produit ou procédé commercial n'implique de la part du Bureau international du Travail aucune appréciation favorable ou défavorable.

Les publications du Bureau international du Travail peuvent être obtenues dans les principales librairies ou auprès des bureaux locaux du BIT. On peut aussi se les procurer directement, de même qu'un catalogue ou une liste des nouvelles publications, à l'adresse suivante: Publications du BIT, Bureau international du Travail, CH-1211 Genève 22, Suisse, ou par courriel: pubvente@ilo.org ou visitez notre site Web: www.ilo.org/publns.

Visitez notre site Web: www.ilo.org/ipec

Imprimé en Italie
Photocomposition par Centre international de formation de l'OIT à Turin, Italie

Remerciements

Je souhaite remercier plusieurs personnes, en particulier:

Mustafa Hakki Ozel pour le suivi permanent et le soutien qu'il a apportés à ce travail; Farhad Mehran, Debi Mondal, Angel Rodriguez, Giulio Ghellini, Gianni Betti et l'ensemble de l'équipe SIMPOC pour la révision du document; les participants au séminaire de révision organisé en mai 2008 au siège de l'OIT à Genève; tous les collègues de l'OIT qui ont discuté des diverses versions du document et formulé des suggestions à cet égard; et Giulia Ciampalini et Ilaria Vannini, de l'Université de Sienne, pour m'avoir aidée à préparer ce manuel.

Vijay Verma

Sienne, le 30 juin 2008.

Sommaire

Remerciements iii

CHAPITRE 1
Introduction 1

1.1. Les sources d'information sur le travail des enfants 1
1.2. Les autres sources de données sur le travail des enfants 3
1.3. La complémentarité des sources 5
1.4. Les enquêtes auprès des ménages 6
1.5. Le contenu de ce manuel 8

CHAPITRE 2
Choisir une structure d'enquête appropriée 11

2.1. Une distinction fondamentale: enquête sur le travail des enfants (ETE) versus enquête sur les enfants travailleurs (EET) 13
2.2. Les relations de l'ETE avec l'enquête de base 24
2.3. L'implication pour l'échantillonnage de la définition du «travail des enfants» utilisée 32
2.4. L'EET intégrée 38
2.5. L'EET liée à l'ETE 40
2.6. Les enquêtes sur les enfants et les enquêtes sur les activités des enfants 45
2.7. Les enquêtes sur les unités autres que les ménages 55
2.8. Enquête sur les activités des jeunes en Afrique du Sud 1999: description et commentaires 56
2.9. Enquête sur l'activité des jeunes en Jamaïque 2002: description et commentaires 70

CHAPITRE 3
L'échantillonnage en vue d'une enquête typique basée sur la population 79

3.1. Introduction 79
3.2. Le plan d'échantillonnage: orientation pratique 79
3.3. La population de l'enquête: la base de sondage 85
3.4. Les écarts par rapport au sondage aléatoire simple: stratification, échantillonnage par grappes et probabilités inégales 90
3.5. Le choix de la taille d'échantillon 100
3.6. L'échantillonnage des unités de zone avec probabilité proportionnelle à la taille (PPT) 114
3.7. L'échantillonnage systématique 115
3.8. Les mesures de taille imparfaites 117
3.9. Comment traiter les très grandes unités 123
3.10. Comment traiter les très petites unités 125

3.11. Une illustration chiffrée 127
3.12. Les tirages non probabilistes 132

CHAPITRE 4
L'enquête sur le travail des enfants (ETE): l'échantillon intégré de taille réduite 143

4.1. La structure de l'ETE et ses liens 145
4.2. La sélection des zones 148
4.3. La distribution de l'échantillon et les domaines de rapport 150
4.4. Réajuster l'échelle des mesures de taille pour faciliter la sélection de l'échantillon 153
4.5. Traiter de «très grandes» zones dans le sous-échantillonnage 157
4.6. De l'EMO à l'ETE: sous-échantillonnage des unités finales au sein des zones de sondage 159
4.7. Traiter de «très petites» zones dans le sous-échantillonnage 161

CHAPITRE 5
L'enquête sur les enfants travailleurs (EET) 165

5.1. L'approche de l'enquête 165
5.2. Sélection des zones 169
5.3. Traiter les «très petites» et les «très grandes» zones 171
5.4. Augmenter la taille des premières zones d'échantillon 174
5.5. L'échantillonnage adapté par grappes 175
5.6. Illustrations numériques 179

CHAPITRE 6
Illustration du plan d'échantillonage et des procédures de sélection 191

6.1. Introduction 191
6.2. Contexte et objectifs de l'enquête dans l'illustration 193
6.3. Echantillonnage pour listage et ETE 196
6.4. Echantillonnage pour l'enquête sur les enfants travailleurs (EET) 207
6.5. Echantillonnage pour le module EMO 215
6.6. Illustration numérique des procédures d'échantillonnage 217
6.7. Illustration de la sélection d'échantillon (échantillonnage systématique circulaire PPT) 224

CHAPITRE 7
Estimation fondée sur les données d'échantillon 243

7.1. Pondération des données d'échantillon 243
7.2. Calcul des poids d'échantillon: une approche systématique 247
7.3. Estimation des ratios et des totaux 254
7.4. Commentaire sur l'estimation d'une petite zone 257
7.5. Importance des informations sur les erreurs d'échantillonnage 264

7.6. Procédures pratiques pour le calcul des erreurs d'échantillonnage 265
7.7. Application des méthodes dans la pratique....276
7.8. Mesures transportables de l'erreur d'échantillonnage 277
7.9. Quelle doit être l'ampleur de l'échantillon pris par grappe? 280

Références....283

Encadrés

Encadré 2.1: Objectifs de l'enquête et population cible. Portugal 1998....16
Encadré 2.2: Exemples de combinaisons EMO-ETE: Costa Rica 2003....29
Encadré 2.3: Exemples de combinaisons EMO-ETE: Turquie 1994 30
Encadré 2.4: Exemple de liaison EMO-ETE: Zimbabwe 1999....31
Encadré 2.5: Implications de la terminologie et des définitions clés pour l'échantillonnage: Portugal 1998. 35
Encadré 2.6: Sur le concept de travail des enfants : Manuel du SIMPOC....36
Encadré 2.7: Exemples d'enquêtes sur les enfants: Ukraine 1999....48
Encadré 2.8: Exemples d'enquêtes sur les enfants: Cambodge 1996 49
Encadré 2.9: Exemples d'enquêtes sur l'activité des enfants: Panama 2000 50
Encadré 2.10: Exemples d'enquêtes sur l'activité des enfants: Belize 2001 50
Encadré 2.11: Exemples d'enquêtes sur l'activité des enfants: Ghana 2001 52
Encadré 2.12: Exemples d'enquêtes sur l'activité des enfants: Mongolie 2002-03....53
Encadré 2.13: Exemples d'enquêtes sur l'activité des enfants: Sri Lanka 1999....54
Encadré 2.14: Exemples d'enquêtes sur l'activité des enfants: République dominicaine 2000.....54
Encadré 2.15: Exemples d'enquêtes sur l'activité des enfants: Costa Rica 2003 55

Tableaux

Tableau 2.1: Enquêtes sur le travail des enfants: de l'enquête de base à l'ETE - Exemples de diversité des structures....22
Tableau 2.2: Enquêtes sur le travail des enfants: de l'ETE à l'EET - Exemples de diversité des structures 23
Tableau 3.1: Quelques illustrations de la variation de la taille d'échantillon selon la proportion estimée....110
Tableau 3.2: L'augmentation requise pour la taille d'échantillon en fonction de l'augmentation de la taille des grappes (b) et de la corrélation intragrappes (roh)....111
Tableau 3.3: Illustration: échantillonnage des unités de zone avec PPT....128
Tableau 4.1: Illustration: sous-échantillonnage de «grandes» zones de l'EMO à l'ETE 158
Tableau 4.2: Équations de sélection supposées de l'EMO et de l'ETE décrivant la procédure de sous-échantillonnage de l'EMO à l'ETE....161
Tableau 4.3: Sous-échantillonnage des zones d'EMO pour l'ETE 164
Tableau 5.1: La sélection de zones à partir de l'échantillon de base....181
Tableau 5.2: La sélection d'unités finales dans les zones sélectionnées....184

Tableau 5.3: Les probabilités de sélection globales et la taille d'échantillon attendue 188
Tableau 6.1: Taille des principaux domaines.....198
Tableau 6.2: Nombre de personnes, de ménages et de ZD en fonction du critère urbain, rural et du gouvernorat (à partir de la base de sondage du recensement).....199
Tableau 6.3: Nombre moyen de ménage par ZD: variation par gouvernorat.....201
Tableau 6.4: Illustration de la répartition de l'échantillon et des taux d'échantillonnage pour la sélection des ZD en fonction du critère urbain/rural et du gouvernorat (Taille d'échantillon cible: 100000 ménages) 205
Tableau 6.5: Caractéristiques sommaires des diverses illustrations de la base de sondage 224
Tableau 6.6: Exemple de procédure d'arrondissement d'une séquence de petits nombres 227
Tableau 6.7: Exemple type de base de sondage des ZD pour l'échantillonnage de la phase 2 227
Tableau 6.8: Exemple d'échantillonnage à deux degrés de la phase 1 à la phase 2.....231
Tableau 6.9: Spécification d'un nombre minimum (e_{min}) de ménages de type E à prélever dans toute zone d'échantillon 234
Tableau 6.10: Limite maximale (e_{max}) du nombre de ménages de type E à prélever dans toute zone d'échantillon 237
Tableau 6.11: Exemple d'échantillonnage systématique circulaire avec une probabilité proportionnelle à la taille 240

Figures

Figure 1: Illustration de la procédure d'échantillonnage systématique.....117
Figure 2: Schéma d'illustration d'une procédure d'estimation de petite zone.....263

Chapitre 1
Introduction

Comme le souligne *Statistiques sur le travail des enfants. Manuel de méthodologies de collecte de données au moyen d'enquêtes* (BIT, 2004, p. 3)[1], le travail des enfants est devenu une question mondiale importante. Des statistiques détaillées et actualisées sur les enfants travailleurs sont nécessaires pour «déterminer l'ampleur et la nature du problème, identifier les facteurs sous-jacents au travail des enfants, révéler ses conséquences et susciter une prise de conscience publique de la constellation liée à ces questions».

Ce manuel est consacré aux questions d'échantillonnage posées lors d'enquêtes ménages sur le travail des enfants. Les considérations qu'implique la conception d'enquêtes visant à estimer l'ampleur et la nature du travail des enfants dans des secteurs et des activités ciblés, ne sont pas évoquées dans ce manuel, ces considérations étant souvent assez différentes de celles intervenant dans les enquêtes conventionnelles, à base plus large, menées auprès des ménages. À certaines fins et dans certaines circonstances, les enquêtes sur le travail des enfants peuvent aussi comporter une collecte de données non basée sur les ménages, voire même contraindre à s'écarter quelque peu des principes de l'échantillonnage probabiliste. Diverses techniques statistiques utilisées pour échantillonner des unités non standardisées doivent être traitées séparément, ce qui pourra faire l'objet de futurs documents du SIMPOC. Néanmoins, un grand nombre des techniques examinées dans ce manuel peuvent aussi être utiles pour la conception d'enquêtes ciblées ou sectorielles, davantage spécialisées sur le travail des enfants.

1.1. Les sources d'information sur le travail des enfants

Les données relatives au travail des enfants peuvent être obtenues à partir de diverses sources, souvent utilisées en combinaison. Elles peuvent inclure des données émanant de recensements de population nationaux, de sources secondaires, d'enquêtes ménages nationales existantes ainsi que d'études et enquêtes spéciales sur le travail des enfants (*Manuel du SIMPOC*, p. 60).

Les recensements

Bien que peu de recensements nationaux de la population fournissent des données sur l'ampleur du travail des enfants, l'information provenant des recensements constitue une base essentielle pour l'interprétation et l'analyse de données sur le travail des enfants émanant d'autres sources. Le recensement de la population est également la source fondamentale d'une base de sondage pour les enquêtes sur le travail des enfants et pour des enquêtes comparables.

---

[1] Auquel il sera désormais fait référence sous l'appellation de *Manuel du SIMPOC.*

Les enquêtes nationales générales auprès des ménages

De nombreux pays recueillent des données socio-économiques et démographiques dans le cadre d'enquêtes ménages périodiques, telles que les enquêtes sur la population active, les conditions de vie, les revenus et les dépenses des ménages, la démographie et la santé, etc. Ces enquêtes ne produisent généralement pas des données détaillées sur le travail des enfants, mais elles peuvent fournir des informations qui seront utiles pour l'analyse de la situation concernant le travail des enfants. De plus, rattacher des modules sur le travail des enfants à ces enquêtes auprès des ménages constitue également une source potentielle d'information.

Les sources secondaires

En ce qui concerne les sources secondaires, le *Manuel du SIMPOC* souligne que toute une série d'institutions qui ne sont pas essentiellement concernées par le travail des enfants apportent souvent des informations utiles dans ce domaine. On en trouve des exemples dans les rapports annuels sur l'école établis par les ministères de l'Éducation, les enquêtes auprès des écoles et les rapports d'inspection, les rapports émanant de bureaux statistiques nationaux, les enquêtes et recherches menées par des organisations internationales pour le développement, et autres études et rapports préparés pour des ministères nationaux et pour la communauté des bailleurs de fonds.

Les études et enquêtes du SIMPOC sur le travail des enfants

Elles comportent plusieurs types d'instruments différents. Le *Manuel du SIMPOC* relève les différentes méthodes suivantes pour la collecte de données sur le travail des enfants:

- les enquêtes ménages sur le travail des enfants;
- les évaluations rapides (ER);
- les enquêtes établissement;
- les enquêtes sur les enfants de la rue;
- les enquêtes auprès des écoles;
- les enquêtes au niveau communautaire; et
- les enquêtes et études de base.

Dans ce manuel d'échantillonnage, nous nous préoccuperons des problèmes d'échantillonnage qui se posent dans le contexte d'enquêtes auprès des ménages sur le travail des enfants ou plus généralement, d'enquêtes portant sur les activités des enfants vivant dans un ménage ordinaire.

Il importe de noter d'emblée le potentiel et les limites de l'information sur le travail des enfants qui peut être recueillie dans le cadre d'enquêtes auprès des ménages. Avant d'exposer le rôle des enquêtes sur le travail des enfants, nous résumerons d'abord les caractéristiques des autres sources d'information. Ces différentes sources sont généralement complémentaires. Les notes ci-dessous, largement inspirées du *Manuel du SIMPOC,* devraient aider à clarifier le potentiel et les limites des méthodes d'enquête auprès des ménages.

1.2. Les autres sources de données sur le travail des enfants

Les méthodes suivantes viennent compléter les enquêtes standard auprès des ménages.

Les évaluations rapides

Les évaluations rapides (ER) ont recours à des techniques largement qualitatives employées pour recueillir des informations approfondies sur les pires formes de travail des enfants (PFITE)[2] cachées ou invisibles de différents types, souvent concentrées dans des secteurs géographiques particuliers. La méthodologie de l'ER a été élaborée pour obtenir des informations sur le travail des enfants et les enfants travailleurs dans les types d'activité et de profession les plus dangereux ou malsains. Elle utilise une approche participative pour mener des discussions et des entretiens sur les conditions de vie et de travail d'enfants impliqués dans des activités ou des professions, difficiles à cerner et à caractériser. Une ER peut recourir aux questionnaires structurés ou semi-structurés, à l'observation rapprochée, aux enquêtés clés ou autres personnes informées et à l'information générale dérivée des études de cas et des rapports administratifs.

La méthodologie de l'ER est essentiellement destinée à fournir des informations de manière relativement rapide et peu onéreuse, par exemple utilisables dans la sensibilisation générale aux problèmes ou la conception de projets spécifiques. Ses résultats sont surtout qualitatifs et descriptifs et elle est habituellement limitée à des secteurs géographiques restreints. L'utilité de cet outil est donc généralement limitée lorsqu'il s'agit d'obtenir des estimations quantitatives de l'ampleur du travail des enfants. Comme pour tout type de collecte de données, la valeur des résultats dépend de la qualité et de l'adéquation de la conception de l'étude. L'utilité des informations qualitatives d'une étude d'ER peut être améliorée en la complétant avec une enquête auprès des ménages, dans le secteur étudié.

Les enquêtes par établissement

La plupart des sources d'information font référence à l'offre de travail des enfants, à savoir les enquêtes sur les enfants offrant leur travail. Pour collecter des informations du côté de la demande, soit les bénéficiaires du travail des enfants, des entretiens avec les employeurs ainsi que d'autres enquêtes basées dans les établissements ou sur les lieux de travail sont nécessaires.

Les questionnaires d'enquête par établissement sont normalement administrés sur les lieux de travail – que ce soit en usine ou dans une unité de production à domicile qui utilise des travailleurs rémunérés – et cherche à obtenir des informations concernant l'unité de production et la nature de sa main-d'œuvre, l'accent étant mis sur les enfants travailleurs. Les salaires des enfants et autres avantages du travail, la nature de l'activité, les horaires de travail, les autres conditions de travail, ainsi que les blessures et maladies liées au travail, sont parmi les informations recherchées. Ces données

[2] L'abréviation PFITE doit être clarifiée dans ce manuel. Elle réfère aux pires formes de travail des enfants autres que les travaux dangereux (couramment appelée " pires formes indiscutables de travail des enfants")

peuvent être comparées à d'autres données parallèles concernant les travailleurs adultes. On peut également rechercher des informations relatives à la perception des employeurs des avantages et des inconvénients à utiliser des enfants travailleurs, les méthodes de recrutement utilisées, etc.

Une enquête par établissement peut être fondée sur un échantillon choisi dans une base d'établissements employant probablement des enfants, auquel cas elle vise à obtenir des estimations pour ce type d'établissement. Le plus souvent, ces enquêtes comportent des entretiens menés dans des établissements choisis pour leurs liens avec des enfants travailleurs identifiés grâce à une enquête auprès des ménages sur le travail des enfants. Dans ce dernier cas, l'enquête par établissement fournit des variables complémentaires sur les enfants avec lesquels les établissements sont liés. Elle ne produit pas nécessairement en elle-même un échantillon représentatif de la population totale des établissements employant des enfants.

Les enquêtes sur les enfants de la rue

Les enfants qui vivent dans les rues ou dans des situations extrêmes de même nature sont les plus vulnérables parmi les groupes d'enfants travailleurs. Ils ne peuvent pas faire l'objet d'enquêtes normales auprès des ménages, car celles-ci excluent les personnes sans domicile. La plupart de ces enfants vont d'un point à un autre pendant la journée et dorment hors des bâtiments la nuit, ce qui rend difficiles les enquêtes menées à leur sujet avec des procédures d'échantillonnage normales. Des enquêtes spécialement conçues pour les enfants de la rue sont nécessaires pour recueillir l'information appropriée. Une approche de l'échantillonnage *de convenance* ou *délibérée* est souvent nécessaire, que ce soit dans le choix des zones à étudier que dans la conduite d'entretiens aléatoires avec les enfants concernant leurs conditions de travail, et avec les employeurs du secteur informel employant des enfants.

Les enquêtes auprès des écoles

Les enquêtes auprès des écoles sont nécessaires pour la collecte de statistiques indiquant l' incidence du travail des enfants sur la fréquentation et les résultats scolaires des enfants, ainsi que leur attitude à l'égard de la scolarité. Elles s'appliquent normalement à des enquêtés identifiés comme enfants travailleurs par les enquêtes ménages. Habituellement, outre les enfants travailleurs, cette approche s'applique aussi aux enfants non travailleurs en tant que groupe de contrôle, de préférence originaires des mêmes écoles fréquentées par les enfants travailleurs enquêtés. Des entretiens sont dirigés vers les enfants, les enseignants, les directeurs d'école et les parents. L'enquête peut aussi chercher à évaluer des facteurs liés à l'école, comme la qualité de l'éducation disponible, qui peuvent influencer la probabilité que les enfants s'engagent dans une activité économique[3].

Les enquêtes au niveau communautaire

Les enquêtes au niveau communautaire recueillent habituellement des informations provenant des administrateurs et autres dirigeants de la communauté pour obtenir un

[3] Dans la cadre de la main-d'œuvre, le concept d'*activité économique* inclut la recherche du travail. Toutefois, dans le contexte du travail des enfants, la recherche du travail est enclue. En ce qui concerne les enfants, c'est dans ce contexte que les termes activités économiques sont utilisés dans ce manuel.

profil culturel, démographique et socio-économique de celle-ci, ainsi que d'autres caractéristiques qui peuvent être corrélées au travail des enfants dans la région. Elles peuvent être utiles pour identifier les principales variables liées directement ou indirectement à la présence du travail des enfants dans la communauté. Il est également fréquent que l'on recueille des informations dans le secteur concernant le niveau de revenus des ménages, la pauvreté, les activités économiques majeures, l'emploi saisonnier, l'alphabétisation et la disponibilité des services et infrastructures publiques (écoles, structures médicales, système de transports, eau, électricité, opportunités de loisirs). Comme l'indique le *Manuel du SIMPOC,* les enquêtes au niveau communautaire peuvent «fort bien être conduites comme des recherches indépendantes de collecte de données sur des situations particulières du travail des enfants… toutefois, il est fréquent que les enquêtes au niveau communautaire soient aussi partie intégrante de la méthodologie appliquée dans les ER [évaluations rapides] et les enquêtes de base».

Les enquêtes et études de base

Les enquêtes et études de base sont utiles pour identifier les populations cibles et leurs caractéristiques, et pour analyser les déterminants et les conséquences du travail des enfants dans des secteurs socio-économiques spécifiques. Elles peuvent faire intervenir une ou plusieurs des méthodes de collecte des données disponibles, comme une combinaison de l'enquête par échantillonnage et des techniques de recherche qualitatives (participatives). Ainsi, ces études peuvent utiliser une combinaison de techniques qualitatives et quantitatives pour recueillir des données sur la situation existante avant une intervention ou un projet, essentiellement en tant qu'outil de planification et de mise en œuvre pour des projets identifiant les cibles auxquelles l'intervention s'adresse. Elles peuvent aussi servir au suivi de la progression d'un projet et à l'évaluation de son impact dans des secteurs ou des zones particuliers. Pour les objectifs du programme d'intervention, les enquêtes de base peuvent collecter des données sur la situation initiale (de base) utilisables à chaque étape du cycle du programme: conception, mise en oeuvre, suivi et évaluation d'impact. L'information fournie aide à établir des cibles qui permettent d'apprécier les changements intervenus grâce à des études complémentaires et d'élaborer des systèmes de suivi.

1.3. La complémentarité des sources

Tout en appréciant les limites et les points forts de l'information sur le travail des enfants qui peut être recueillie grâce aux enquêtes ménages, il importe également de noter que les diverses sources d'information sur le travail des enfants sont généralement complémentaires. L'implication pratique du point de vue de l'enquête et de la conception de l'échantillonnage est que l'instrument conçu au niveau des ménages n'a pas besoin d'être développé pour répondre à tous les besoins d'information: en fait, d'autres types d'instruments peuvent mieux répondre à ces besoins – ou même sont seuls en mesure de le faire. Comme l'indique le *Manuel du SIMPOC* (p. 70), «l'expérience a démontré que la collecte de données complètes sur le travail des enfants est une tâche extrêmement difficile, et aucune méthode d'enquête n'est en soi satisfaisante pour les besoins en données».

Cela est dû en premier lieu à la complexité de ces besoins: «Les enfants travaillent dans un large éventail de circonstances, et aucune technique seule ne peut être adaptée à l'enquête de toutes ces situations. En outre, l'analyse de la politique et du projet cible d'intervention exige des informations d'une variété d'enquêtés potentiels qui peuvent avoir une influence sur la vie et le développement d'un enfant. Ce sont les enfants eux-mêmes, les parents ou les tuteurs, les patrons, les enseignants de l'école, les leaders communautaires, les pairs de l'enfant et les frères et sœurs. Les circonstances dans la maison, l'école, le lieu de travail, et la grande communauté à laquelle l'enfant appartient, tout cela a un lien avec les résultats et les caractéristiques du travail des enfants. Collecter toutes les données pertinentes des sources requises au moyen d'une seule enquête ou à une seule occasion est impossible».

En second lieu, nous devons faire face aux problèmes particuliers posés par les pires formes de travail des enfants. Avec les PFTE, il est très difficile, voire impossible, d'entrer en contact avec l'enfant pour recueillir l'information nécessaire. Comme elles demeurent généralement cachées, il n'existe pas de plan d'échantillonnage adéquat pour leur dénombrement. De plus, les échantillons requis ne peuvent être conçus et choisis sans informations préalables sur la localisation, les caractéristiques et la situation des enfants engagés dans les pires formes intrinsèques du travail des enfants (PFITE). En conséquence, les enquêtes régulières auprès des ménages sont largement exclues dans cette perspective; des procédures spéciales d'échantillonnage et de dénombrement doivent être employées.

1.4. Les enquêtes auprès des ménages

Les enquêtes auprès des ménages régulières sont des enquêtes par sondage menées auprès des ménages ayant pour cible les enfants, mais aussi leurs parents/tuteurs vivant sous le même toit. Ces enquêtes peuvent être conduites de manière indépendante, ou en tant que modules rattachés à d'autres enquêtes nationales auprès des ménages, comme les enquêtes sur la population active (EPA). Les statistiques générées par ces enquêtes incluent les activités économiques et non économiques (comme les tâches ménagères) des enfants, les horaires de travail, la nature des travaux effectués, les questions de santé et de sécurité, dont les blessures subies au travail, et aussi des variables générales comme les caractéristiques démographiques et sociales des membres du ménage, et autres caractéristiques élémentaire de celui-ci.

Le *Manuel du SIMPOC* (p. 79-90) note ce qui suit concernant les enquêtes auprès des ménages sur le travail des enfants:

- Le ménage est généralement l'unité la mieux appropriée pour identifier les enfants et leurs familles, apprécier leurs caractéristiques socio-économiques et démographiques et leurs conditions de logement, et obtenir des informations concernant les circonstances qui contraignent les enfants à travailler, et les conditions de travail pour les activités inscrites dans le cadre du ménage.
- Les enquêtes auprès des ménages basées sur un échantillonnage probabiliste offrent une approche efficiente pour estimer l'ampleur de formes particulières de travail des enfants, à l'exception de certaines catégories spéciales de travail des enfants, comme les enfants vivant dans la rue ou ceux qui sont engagés dans des formes cachées de travail des enfants.

- La principale limite vient du fait que les enquêtes auprès des ménages ne sont généralement pas appropriées à la collecte d'informations auprès d'enfants engagés dans les pires formes intrinsèques du travail des enfants (PFITE).
- Dans la mesure où les enquêtes ménages sont basées sur des échantillons scientifiquement conçus, elles permettent la généralisation des résultats de l'étude à la population générale dans laquelle l'échantillon a été prélevé. Pour cela, un échantillon d'enquête sur le travail des enfants devra être choisi conformément à des principes établis: il devra être représentatif de la population tout entière, et l'on devra être en mesure de produire des estimations des paramètres de population dans des marges d'erreur connues et acceptables.
- Dans la pratique, le contenu de l'enquête auprès des ménages peut être détaillé et spécialisé, en fournissant des informations sur la dynamique du travail des enfants ou les flux bruts entre différentes catégories de travail des enfants; ou bien l'enquête peut être concentrée sur quelques caractéristiques de base des enfants travailleurs. Le choix dépend des besoins de données, des ressources disponibles et des dispositions et des circonstances en fonction desquelles l'enquête est menée.
- Dans une enquête sur le travail des enfants, les enquêtés clés sont les enfants travailleurs et potentiellement travailleurs et leurs parents ou tuteurs. Les questionnaires utilisés jusqu'ici dans ces enquêtes sont généralement conçus pour obtenir des informations concernant l'ampleur, les caractéristiques et les conditions du travail des enfants.
- Ces questionnaires recueillent aussi des informations sur les conditions de travail, la branche d'activité, la profession, le statut du travail et les effets sur la santé de l'enfant, ses résultats scolaires et ses possibilités de développement normal. De nombreux sujets liés au travail des enfants peuvent être intégrés dans les questionnaires d'enquête, parmi lesquels les caractéristiques démographiques et socio-économiques, les conditions de logement, les caractéristiques des enfants et de leurs familles en relation avec le travail, les facteurs qui conduisent les enfants au travail, et les attitudes des parents/tuteurs envers le travail et la scolarité des enfants.
- Les enquêtes auprès des ménages peuvent appliquer des conceptions et des structures organisationnelles diverses. Les principaux facteurs qui déterminent la conception sont évidemment des objectifs de fond concernant le contenu, la complexité et la périodicité de l'information recherchée. Ces besoins fondamentaux déterminent des caractéristiques de la structure de l'enquête comme son calendrier, sa fréquence, la période de référence et les dispositions relatives à l'échantillonnage. Ainsi, par exemple, il peut s'agir d'une enquête continue (habituellement conçue pour obtenir des données régulièrement échelonnées dans le temps sur les niveaux et les tendances du moment), ou d'une enquête occasionnelle (normalement conçue pour obtenir des informations de référence et structurelles).

La plupart des enquêtes récentes sur le travail des enfants appuyées par SIMPOC ont été des enquêtes occasionnelles ou uniques.

1.5. Le contenu de ce manuel

Les enquêtes auprès des ménages sur le travail des enfants peuvent être de types divers et comporter des types différents de structure et de conception pour refléter des objectifs et des situations distincts. La conception de l'échantillon d'une enquête ménages sur le travail des enfants commence par le choix d'une structure d'enquête appropriée. Cette structure renvoie à la manière dont ses différentes composantes seront organisées les unes par rapport aux autres, et le cas échéant, par rapport à d'autres enquêtes.

Le **chapitre 2** présente une typologie des structures d'enquête et examine les facteurs sous-jacents au choix d'une structure appropriée. Un grand nombre d'enquêtes sur le travail des enfants dans les pays en développement se présentent sous la forme de tableaux et comportent une analyse du type de liaison rencontré entre les «enquêtes de base» et les enquêtes visant à estimer l'ampleur du travail des enfants, et entre ces dernières et des enquêtes plus détaillées sur la situation des enfants travailleurs. Eu égard au fait qu'une condition préalable absolument fondamentale d'une bonne conception de l'échantillon est que «*l'échantillonneur doit comprendre le contenu de l'enquête pour laquelle l'échantillon doit être conçu, les objectifs et la méthodologie de l'enquête et les conditions de sa mise en œuvre»,* le but de ce chapitre est d'en donner une idée, à partir de la pratique réelle, pour les enquêtes sur le travail des enfants.

Le **chapitre 3** vise à clarifier les principes et procédures de base de l'échantillonnage intervenant dans les types de conception les plus couramment utilisés dans les enquêtes ménages qui concernent l'ensemble de la population, auxquelles peut être rattachée une enquête sur le travail des enfants, sous la forme d'un module ou par tout autre moyen de liaison. Bien entendu, ceux-ci sont directement appropriés à la conception d'enquêtes indépendantes sur le travail des enfants. Cette description des principes et des procédures de base de l'échantillonnage est importante pour mieux cerner les différentes dispositions pour l'appréciation du travail des enfants pouvant être dérivées ou liées à d'autres enquêtes auprès des ménages. Ce chapitre traite un certain nombre d'aspects pratiques importants concernant les bases de sondage et la conception des échantillons, les méthodes de sélection, puis assez longuement, la question complexe de la taille des échantillons.

Les enquêtes visant à estimer l'ampleur du travail des enfants tendent à poser des conditions en matière d'échantillonnage semblables à celles des enquêtes sur la population générale, telles que les enquêtes sur la main-d'œuvre. Cela vient du fait que les populations de base pour les deux types d'enquête (tous les enfants pour les enquêtes sur le travail des enfants; la population adulte pour les enquêtes sur la main-d'œuvre) tendent à être réparties de manière similaire.

Le **chapitre 4** décrit les détails techniques des procédures d'établissement de l'échantillon d'une enquête sur le travail des enfants (ETE), sur la base de l'échantillon utilisé pour une enquête plus vaste sur la population générale, telle que l'EPA.

Pour des informations détaillées sur les conditions et les conséquences du travail des enfants, la population de base qui nous intéresse est la population d'enfants travailleurs, plutôt que celle de l'ensemble des enfants: nous appelons les enquêtes menées sur cette

base «enquêtes sur les enfants travailleurs» (EET). Les conditions d'échantillonnage de l'EET diffèrent de celles qui s'appliquent à une enquête sur la population générale telles que l'EPA, comme de celles concernant une enquête sur la population générale des enfants telles que l'ETE. Cela résulte des différences existant entre l'EET et ces autres types d'enquête en ce qui concerne la distribution des populations de base respectives. La distinction entre les deux types d'enquête - ETE et EET, selon la terminologie utilisée dans ce manuel - est fondamentale. Cela ne signifie pas que l'ETE et l'EET doivent être deux enquêtes distinctes: elles peuvent constituer des composantes d'une seule et même enquête – et il en est souvent ainsi. Toutefois, même quand ces deux types d'enquête doivent être intégrés dans une seule et même opération, la distinction demeure conceptuellement importante, car cela permet de choisir de manière appropriée la conception de compromis de l'opération intégrée.

Le **chapitre 5** examine les questions de conception d'un échantillon pour une enquête portant sur la population de base des enfants travailleurs et expose les détails techniques des procédures permettant d'établir un échantillon d'EET sur la base d'une enquête plus vaste portant sur la population générale, comme l'EPA ou l'ETE. La condition fondamentale est qu'une enquête intensive visant à étudier en détail la situation et les conditions de vie des enfants travailleurs doit avoir une structure d'échantillon qui reflète le modèle de distribution de ces enfants dans la population générale.

Le **chapitre 6** vise à élucider le processus d'élaboration des procédures de conception et de sélection de l'échantillon dans une situation réelle. Dans ce chapitre est présentée une illustration très réaliste élaborée dans le contexte d'une enquête réelle menée dans un pays en développement. Elle comporte de nombreux détails et exemples chiffrés destinés à clarifier les procédures dans leur diversité.

La structure d'enquête élaborée et présentée dans ce chapitre diffère quelque peu du modèle plus courant qui, est sous-jacent aux deux chapitres précédents. Il n'existe pas de conception ou de structure figées pour les enquêtes sur le travail des enfants. Les conditions à remplir en matière de données sont diverses et doivent être établies et adaptées de la manière la plus efficiente possible.

Enfin, le **chapitre 7** traite une sélection de problèmes relatifs aux estimations effectuées à partir de données d'enquêtes par sondage dans une perspective pratique. Il étudie les questions et les procédures statistiques relatives à la pondération des données d'échantillon et à la production d'estimations à partir des enquêtes. Il examine ensuite les procédures pratiques de calcul et d'analyse de l'information sur les erreurs d'échantillonnage, avant de tirer en conclusion quelques leçons concernant le choix de la taille et de la structure de l'échantillon d'un point de vue pratique.

Chapitre 2
Choisir une structure d'enquête appropriée

Les enquêtes ménages sur le travail des enfants peuvent être de types divers et comporter des types différents de structure et de conception reflétant des objectifs et des situations distincts.

Dans la conception de l'échantillon d'une enquête ménage sur le travail des enfants, le premier pas consiste à choisir une structure d'enquête appropriée. Une enquête sur le travail des enfants doit collecter différents types d'information, depuis l'estimation de l'ampleur du travail des enfants jusqu'à l'étude approfondie des conséquences du travail sur les enfants. Des types de données varient par leur contenu, leur mode de collecte et – ce qui s'applique particulièrement à l'échantillonnage – la population de base concernée.

Les différents types de données impliqués peuvent être considérés comme constituant différentes «composantes» de l'enquête. Par «structure d'enquête», nous entendons la manière dont différentes composantes de l'enquête ont été organisées les unes par rapport aux autres. Ainsi, par exemple, ces composantes peuvent être combinées en un seul ensemble intégré, ou bien elles peuvent rester distinctes mais liées de différentes manières; ou encore elles peuvent constituer des opérations «indépendantes» plus ou moins séparées, chacune constituant comme une enquête en elle-même. De même, l'enquête sur le travail des enfants ou ses composantes peuvent être reliées de diverses manières à d'autres enquêtes existantes, telle que l'enquête sur la main-d'œuvre.

Bien entendu, les différents types d'enquête sur le travail des enfants ont beaucoup de points communs, mais il peut aussi exister entre eux des différences majeures, et il est très utile de discerner et de classer ces différences.

Une typologie des enquêtes clairement établie peut aider à déterminer le type de structure et de conception de l'enquête nécessaire pour répondre aux objectifs fixés et aux contraintes pratiques que l'on rencontrera en entreprenant l'enquête. On trouve dans l'expérience de certains pays de nombreux exemples d'enquêtes sur le travail des enfants dans lesquelles un défaut de compréhension du type d'enquête requis a amené le choix d'une structure, d'une conception ou d'une taille d'échantillon inappropriées à l'enquête.

Une terminologie clairement énoncée est aussi un outil utile, dans la mesure où elle fournit un cadre permettant de décrire et de documenter la méthodologie et la conception de l'enquête de manière claire et détaillée, et de débattre des mérites et des faiblesses relatifs des différents choix.

Nous proposerons dans ce chapitre un grand nombre d'exemples afin d'illustrer des aspects divers de la structure des enquêtes sur le travail des enfants, ainsi que de leurs relations mutuelles ou avec des enquêtes menées sur d'autres thèmes. Comme nous l'avons mentionné en introduction, une condition préalable fondamentale pour la bonne conception d'une enquête est que l'échantillonneur *doit comprendre le contenu de l'enquête pour laquelle l'échantillon doit être conçu, les objectifs et la méthodologie*

de l'enquête et les conditions de sa mise en œuvre. Le but de ce chapitre est d'illustrer ces questions pour les enquêtes sur le travail des enfants à partir de la pratique réelle.

La plupart des exemples proposés dans ce chapitre sont des extraits de rapports nationaux sur des enquêtes consacrées au travail des enfants, qui sont cités sans avoir été révisés. Des commentaires mettant en exergue les points particulièrement intéressants qui se dégagent de ces illustrations les accompagnent dans le corps principal du texte.

On trouvera ci-dessous la liste des encadrés figurant dans ce manuel:

Encadré 2.1:	Objectifs de l'enquête et population cible: Portugal 1998
Encadré 2.2:	Exemples de «combinaisons» EMO-ETE: Costa Rica 2003
Encadré 2.3:	Exemples de «combinaisons EMO-ETE»: Turquie 1994
Encadré 2.4:	Exemple de «liaison EMO-ETE». Zimbabwe 1999
Encadré 2.5:	Implications des définitions clés pour l'échantillonnage: Portugal 1998
Encadré 2.6:	Sur le concept de travail des enfants: Manuel du SIMPOC (BIT, 2004)
Encadré 2.7:	Exemples d'enquêtes sur les enfants: Ukraine 1999
Encadré 2.8:	Exemples d'enquêtes sur les enfants: Cambodge 1996
Encadré 2.9:	Exemples d'enquêtes sur l'activité des enfants: Panama 2002
Encadré 2.10:	Exemples d'enquêtes sur l'activité des enfants: Belize 2001
Encadré 2.11:	Exemples d'enquêtes sur l'activité des enfants: Ghana 2001
Encadré 2.12:	Exemples d'enquêtes sur l'activité des enfants: Mongolie 2002-03
Encadré 2.13:	Exemples d'enquêtes sur l'activité des enfants: Sri Lanka 1999
Encadré 2.14:	Exemples d'enquêtes sur l'activité des enfants: République dominicaine 2000
Encadré 2.15:	Exemples d'enquêtes sur l'activité des enfants: Costa Rica 2003

A la fin de ce chapitre (sections 2.8 et 2.9) des études plus détaillées seront fournies, accompagnées de commentaires, sur les aspects de la conception de deux enquêtes concernant le travail des enfants: l'Enquête sur l'activité des jeunes en Afrique du Sud (1999) et l'Enquête sur l'activité des jeunes de la Jamaïque (2002). L'objectif est d'utiliser ces exemples pour illustrer et examiner de manière critique un certain nombre de points importants en rapport avec la structure et l'échantillonnage des enquêtes.

Dans les tableaux 2.1 et 2.2 qui figurent à la fin de la section 2.1, nous avons essayé de classer, en utilisant la terminologie décrite dans ce chapitre, les structures d'enquête utilisées dans quelques 30 enquêtes sur le travail des enfants menées au cours de ces dernières années. En différents points de ce chapitre, nous commenterons cette diversité des structures d'enquête afin d'évoquer les problèmes importants de la conception des échantillons.

2.1. Une distinction fondamentale: enquête sur le travail des enfants (ETE) versus enquête sur les enfants travailleurs (EET)

2.1.1. L'enquête sur le travail des enfants (ETE)

Nous utilisons l'expression générique «enquête (régulière auprès des ménages) sur le travail des enfants» pour indiquer une enquête ménage par sondage qui a pour objectif essentiel de fournir des informations sur le phénomène du travail des enfants – son ampleur, sa distribution, ses formes, ses secteurs économiques, etc., ainsi que ses conditions, ses caractéristiques et ses conséquences. Le terme «régulière» est utilisé, le cas échéant, pour souligner que le contexte est celui d'une large enquête auprès des ménages, distincte en cela des autres types d'enquête concernant des enfants qui ne résident pas à des domiciles privés – ou du moins qui ne sont pas identifiés pour être dénombrés dans ce contexte.

Tout en maintenant l'utilisation descriptive plus générale de l'expression «enquêtes sur le travail des enfants», il est très utile de garder présents à l'esprit deux types différents de ces enquêtes: elles diffèrent par leurs objectifs, ou du moins par l'accent qui est mis sur différents types d'objectifs.

Le premier type est celui des enquêtes où l'objectif premier est de *mesurer l'ampleur du travail des enfants*. Les enquêtes peuvent aussi étudier les variations de cette ampleur selon la localisation géographique, le type de lieux (urbain-rural), le type et les caractéristiques des ménages, l'emploi et le niveau de revenus des ménages, l'âge et le sexe des enfants, et autres facteurs semblables.

La population cible d'une enquête ayant ce type d'objectifs est la *population totale d'enfants* exposés au risque de travail des enfants. Cette population de base se définit essentiellement en termes de limites d'âge et tend donc à être bien distribuée dans la population générale. La taille et la structure de l'échantillon sont largement déterminées par la taille et la distribution de la population de tous les enfants, ou plus couramment par son approximation: la taille et la distribution de la population générale.

Nous proposons de limiter l'expression «enquête sur le travail des enfants» (ETE) à des enquêtes dont l'objectif premier est de mesurer l'ampleur du travail des enfants, à distinguer de l'«enquête sur les enfants travailleurs» (EET) décrite ci-dessous. Cette distinction a d'abord été introduite dans le *Manuel du SIMPOC* (p. 194). Le facteur déterminant de cette distinction est la population de base pour laquelle sont générées les estimations de l'enquête – pour l'essentiel, tous les enfants dans des limites d'âge spécifiées pour l'ETE, et seulement ceux qui sont considérés comme étant engagés dans le travail des enfants pour l'EET.

2.1.2. L'enquête sur les enfants travailleurs (EET)

Nous avons affaire à un type d'enquête différent quand l'objectif premier est d'enquêter sur les circonstances, les caractéristiques et les conséquences du travail des enfants: les types d'enfants qui sont engagés dans des activités liées au travail, les types de travail

effectué par les enfants, les circonstances et les conditions dans lesquelles travaillent les enfants, l'incidence du travail sur leur éducation, leur santé, leur développement physique et moral, et ainsi de suite. Les objectifs peuvent aussi inclure des recherches sur les causes et les conséquences immédiates de l'entrée des enfants dans le travail. Nous parlons alors d'une «enquête sur les enfants travailleurs» (EET).

La population de base appropriée à l'EET est la *population des enfants travailleurs.* Le concept d'EET signifie que, lorsque l'objectif est de déterminer les conditions et les conséquences du travail des enfants, et non plus son ampleur parmi la totalité des enfants, il serait alors approprié que *la taille et la structure de l'échantillon soient largement déterminées par la taille et la distribution de la population d'enfants travailleurs*.

Il importe en même temps de préciser clairement que le concept d'«enquête sur les enfants travailleurs» n'implique pas que les unités finales dénombrées dans l'enquête soient seulement les enfants travailleurs. Au contraire, il sera normalement nécessaire dans ce type d'enquête de dénombrer *les groupes comparables d'enfants qui ne sont pas engagés dans le travail*, afin de disposer d'un groupe de contrôle pour établir des comparaisons avec les caractéristiques et la situation de ceux qui sont soumis au travail des enfants. Néanmoins, la taille de l'échantillon et la conception d'une EET sont déterminées en premier lieu par le besoin de représenter la population des enfants travailleurs; tout échantillon d'enfants non travailleurs vient en supplément et est choisi et ajouté à l'échantillon principal selon les nécessités de l'analyse.

2.1.3. Les implications pour l'échantillonnage

Il existe certaines différences importantes concernant les aspects de l'échantillonnage entre les enquêtes sur le travail des enfants et les enquêtes sur les enfants travailleurs dans l'acception exposée ci-dessus.

La population cible de l'ETE, composée de la totalité des enfants d'une certaine catégorie d'âge, tend à être distribuée de manière très semblable à la distribution de la population générale. Par conséquent, la structure et la distribution requises pour l'échantillon d'ETE seront probablement très semblables à celles d'une enquête sur la population générale, notamment celles de l'enquête sur la main-d'œuvre (EMO), très comparable à l'ETE dans ses concepts, ses définitions et même son contenu.

La population cible de l'EET – qu'il s'agisse de la population d'enfants engagés dans une activité liée au travail, ou qu'elle soit définie plus étroitement comme constituée par les enfants engagés dans des formes spécifiques de travail des enfants – est, par comparaison, plus restreinte et plus inégalement répartie, souvent concentrée dans des secteurs particuliers. Généralement, la conception de l'échantillon requis est donc également différente de celle de l'EMO ou de l'ETE.

Les deux types d'enquête diffèrent par leur taille et leur complexité. L'ETE est normalement moins intensive (autrement dit, elle implique un entretien d'enquête plus simple et plus bref), et elle exige des échantillons de plus grande taille. La première considération statistique à prendre en compte est la précision avec laquelle doit être estimée la proportion d'enfants engagés dans le travail des enfants, ainsi que les domaines de rapports qui exigent des estimations distinctes.

Par contre, pour les recherches approfondies sur les conditions et les conséquences du travail des enfants, l'EET implique une collecte de données plus intensive, comportant souvent des entretiens séparés avec les tuteurs comme avec les enfants concernés, la collecte de données attitudinales et autres et la conduite d'enquêtes associées, menées par exemple à l'école ou sur les lieux de travail des enfants de l'échantillon. En conséquence, la taille appropriée d'un échantillon d'EET sera probablement beaucoup plus réduite que dans le cas d'une ETE menée dans des circonstances semblables. Pour une enquête intensive comme une EET, les échantillons de grande taille sont souvent inutiles d'un point de vue statistique et sont de toute manière exclus pour des raisons pratiques et économiques. En fait, utiliser un échantillon «trop grand» dans une enquête intensive peut souvent affecter la qualité et la valeur de l'information collectée, dans la mesure où cela fait obstacle à un contrôle étroit de l'opération (voir la section 3.5 sur le choix de la taille des échantillons).

Ce qui précède ne signifie pas que l'ETE et l'EET doivent toujours être deux enquêtes séparées. Au contraire, comme il est indiqué dans la section 2.2 ci-dessous, les deux composantes (ETE et EET) identifiées ici sont, en fait, souvent intégrées sur le plan opérationnel en une seule enquête. Cela ne dispense pas d'examiner les conditions liées à chaque composante en matière de taille et de conception des échantillons: la conception commune de l'enquête intégrée devra être un compromis approprié entre ces conditions.

2.1.4. La diversité des populations cibles: Une illustration

Dans la pratique, les populations cibles intéressant l'enquête peuvent être plus diverses que la distinction entre l'ensemble des enfants et les enfants travailleurs établie par l'ETE et l'EET. La diversité des objectifs d'une enquête sur le travail des enfants et des groupes de population auxquels s'appliquent les résultats de l'enquête est bien illustrée par la description suivante, provenant du Portugal (1998). Selon les questions posées, cette enquête détermine sept groupes cibles différents (voir l'encadré 2.1 ci-dessous pour plus de détails).

Pour trois de ces groupes, à savoir:

Groupe 1: les familles et les activités des enfants,

Groupe 2: les enfants et les activités en général, et

Groupe 7: les aspects de la vie des enfants,

la population cible est composée de tous les enfants (et de leurs familles). Les modules du questionnaire qui s'y rapportent constituent la composante «ETE» de l'enquête. Le groupe 1 est la base d'estimation de la proportion d'enfants engagés dans le travail des enfants.

Pour trois autres groupes, à savoir:

Groupe 3: les enfants ayant une activité économique,

Groupe 5: la caractérisation de ceux qui sont responsables des enfants ayant une activité économique, et

Groupe 6: les attitudes et les perceptions de l'enfant et des adultes qui sont responsables de celui-ci concernant le travail des enfants,

la population cible est constituée par les enfants travailleurs, définis ici comme étant ceux qui sont engagés dans n'importe quelle activité économique. Les modules du questionnaire qui s'y rapportent constituent la composante «EET» de l'enquête. Le groupe 3 – enfants engagés dans une activité économique – est la base; la population des personnes responsables des enfants dans les groupes 5 et 6 est définie par association avec la population de base du groupe 3.

Groupe 4: les enfants qui effectuent des tâches domestiques couvre un groupe d'enfants quelque peu différent. On considère souvent qu'il est moins important que le groupe 3 dans la détermination des besoins de l'EET en matière de taille de l'échantillon[4].

La taille finale de l'échantillon est généralement un compromis entre les conditions des composantes «ETE» et «EET». En principe, il est bien entendu possible d'introduire un sous-échantillonnage du groupe 1 au groupe 3 (autrement dit de suivre seulement un sous-échantillon des enfants travailleurs identifiés dans le groupe 1); ou bien d'introduire un sous-échantillonnage du groupe 3 aux groupes 5-6 (autrement dit de suivre seulement un sous-échantillon des adultes responsables des enfants travailleurs). Une plus grande flexibilité est assurée à cet égard quand les diverses composantes impliquées sont séparables l'une de l'autre sur le plan opérationnel.

Encadré 2.1: Objectifs de l'enquête et population cible. Portugal 1998

Objectif de l'enquête

Parmi la série d'enquêtes déjà adoptées par l'OIT pour étudier le travail des enfants qui ont été mentionnées précédemment, il a été proposé de mener au Portugal trois types d'opérations: un questionnaire destiné aux familles; un questionnaire destiné aux écoles (enseignants et élèves); la collecte d'information continue auprès d'un groupe de personnalités à définir (informateurs clés). L'objectif général du questionnaire destiné aux familles était de quantifier le travail des enfants au Portugal à un niveau global et régional et de caractériser les facteurs déterminants et explicatifs du phénomène. Cet objectif général peut être scindé en deux objectifs spécifiques:

- quantifier le travail des enfants conformément aux recommandations de l'OIT, à la législation portugaise, au concept de travail dans son acception économique et dans un sens plus large, et en termes de temps au cours de l'année précédente et au moment de l'administration du questionnaire;
- caractériser le travail des enfants en termes de facteurs expliquant son existence, de facteurs déterminants dans le contexte social et familial et de positions des familles et des enfants concernant le travail. ….

Questionnaire d'enquête

Selon le modèle des questionnaires corrélés administrés par l'OIT, le questionnaire a été divisé en deux modules.

[4] Des débats considérables sont en cours sur la définition de ce qui constitue le «travail des enfants», et notamment sur la question de savoir si les tâches domestiques «substantielles» devraient y être incluses. Voir *Manuel du SIMPOC,* chapitre 2, ainsi que l'encadré 2.6 ci-dessous.

1. Le premier, qui devait être rempli par le représentant du groupe familial, contenait trois types de questions: sur la famille en général (ex.: logement, mobilier, revenu familial…), sur chacun des membres du groupe familial (sexe, âge, qualification, situation de travail), et sur les enfants selon les paramètres de cette étude (relations entre les responsables et les enfants, activités des enfants…).

2. Les questions du second module, auxquelles les enfants ont répondu eux-mêmes, portaient sur des aspects de la caractérisation, la scolarité, la relation avec l'école, l'occupation du temps libre et les habitudes. Dans le cas de ceux qui participaient à des activités, un second type de questionnaire concernait le caractère de l'activité, ses relations avec le travail, les horaires et les sentiments de l'enfant par rapport à l'activité. …

Groupes de population cibles

… La population cible choisie incluait les enfants de 6 ans, âge auquel les enfants entrent généralement à l'école primaire, à 15 ans, qui est normalement l'âge de la fin de la scolarité obligatoire: c'est aussi la limite légale pour l'admission dans le monde du travail, mis à part certaines situations exceptionnelles et dans le cadre de certaines limites. …

Parmi les objectifs visés dans l'élaboration de ce questionnaire figure l'obtention, à partir des résultats, d'une réponse détaillée à différents groupes de questions.

Groupe 1: Les familles et les activités des enfants

Ce groupe de questions a été établi pour savoir dans combien de familles existe le travail des enfants et d'en connaître la composition socio-économique en relation avec la vie et l'activité des enfants. Et aussi de déterminer s'il existe une corrélation simple entre le niveau de vie, le travail des enfants et d'autres activités des enfants, ou s'il faut compter avec l'intervention d'autres aspects, comme les facteurs régionaux, sociaux et culturels.

Groupe 2: Les enfants et les activités en général

Sans englober l'activité économique au sens strict, le questionnaire vise à explorer les occupations des enfants en un sens plus large, à savoir dans le domaine de l'école, du foyer, du travail au sens économique du terme, des tâches domestiques et du temps libre. Et à faire comprendre la relation existant entre les différentes activités des enfants et le sexe, l'âge, la scolarité et les régions où ils vivent. Si le point de référence essentiel du questionnaire est la semaine précédant son administration, on cherche également à collecter des données relatives aux activités de périodes de vacances et à l'année précédente.

Groupe 3: Les enfants ayant une activité économique

Les enfants engagés dans une activité économique constituent, au sens du questionnaire, le groupe de travail des enfants. Tout d'abord, il convient toutefois de distinguer deux sous-groupes importants pour caractériser le travail

des enfants: les enfants qui travaillent dans des unités économiques à gestion familiale et ceux qui travaillent pour un tiers à l'extérieur de la famille. Dans le cadre de cette distinction, on a cherché à obtenir un ensemble de caractéristiques telles que les régions, le sexe, l'âge, la scolarité et leurs relations au travail. Au sein de ce groupe de questions, un autre objectif a consisté à caractériser le travail des enfants en termes de secteurs d'activité, de types d'unité économique où travaillent les enfants, de raisons de mener ces activités, d'horaires, de rémunération (pour ceux qui travaillent pour un tiers), tout en examinant également quelques données relatives aux maladies et aux accidents provoqués par le travail des enfants.

Groupe 4: Les enfants qui effectuent des tâches domestiques

Les enfants qui effectuent des tâches domestiques ne rentrent pas dans le cadre du travail des enfants. Toutefois, ce type d'activité, quand il dépasse certaines limites, peut lui aussi être préjudiciable à la scolarité et à d'autres aspects précédemment mentionnés. Par conséquent, comme pour le travail des enfants, les enfants qui ont des activités domestiques sont également quantifiés et caractérisés de manière semblable à ceux qui exercent une activité économique. La durée et le type de ces occupations, ainsi que leurs relations avec d'autres aspects comme la scolarité et le temps libre, constituent aussi un axe de recherche de ce questionnaire.

Groupe 5: La caractérisation de ceux qui sont responsables des enfants ayant une activité économique

Avec ces résultats, on s'efforce de connaître les parents (ou les responsables) et d'établir des relations avec l'activité économique des enfants: âge auquel ils ont commencé à travailler, niveau de scolarité atteint, situation professionnelle et niveau de rémunération.

Groupe 6: Les attitudes et les perceptions de l'enfant et des adultes qui sont responsables de celui-ci concernant le travail des enfants

Concernant les enfants qui travaillent, on tente de savoir ce que ressentent les parents, comment ils justifient cette situation, et comment les enfants eux-mêmes acceptent et approuvent le fait de travailler.

Groupe 7: Les aspects de la vie des enfants

Le questionnaire visait aussi à découvrir d'autres aspects de la vie des enfants, en rapport ou non avec leurs activités. Les occupations du temps libre et des périodes de vacances, les relations avec d'autres enfants et chefs de famille, l'heure à laquelle ils se couchent et se lèvent habituellement, le nombre d'heures passées à regarder la télévision, le fait de recevoir ou non de l'argent de poche, ainsi que son montant, sont, entre autres, l'objet de questions destinées à découvrir certains aspects de la vie des enfants qui sont traités dans cette étude.

Source: Child Labour in Portugal: Social Characterisation of School Age Children and their Families, 1998. Portugal Ministère du Travail et de la Solidarité (MTS).

2.1.5. Exemples de diversité structurelle des enquêtes sur le travail des enfants

Les tableaux 2.1 et 2.2 fournissent quelques informations essentielles sur les échantillons et la structure des liaisons de quelques 30 enquêtes sur le travail des enfants. Ils montrent la diversité des structures d'enquête que l'on rencontre. Les concepts relatifs à la structure des enquêtes sont exposés plus en détail dans d'autres sections de ce chapitre.

Liaison de l'ETE à une enquête de base

Le tableau 2.1 classe les types de liaison entre l'ETE et son enquête de base, si elle existe. Si l'ETE n'est pas liée à une enquête de base, elle est classée comme étant «indépendante».

Comme nous l'avons résumé ci-dessus, les deux formes d'arrangements «extrêmes» ont dominé dans les enquêtes passées. La moitié des enquête étudiées (15 sur 29) sont *indépendantes*, ce qui veut dire que l'enquête porte exclusivement ou essentiellement sur le travail des enfants, si bien que son contenu, sa conception et ses procédures peuvent être choisis de manière indépendante (section 2.2.2).

L'autre disposition couramment utilisée consiste à collecter l'information de l'enquête de base et de l'ETE au cours de la même opération (tel est le cas de 13 des 29 enquêtes étudiées). Nous établissons ici une distinction entre les *enquêtes modulaires* sur le travail des enfants, qui viennent essentiellement s'ajouter à un entretien d'enquête de base existante sous la forme d'un module (section 2.2.1), et les *enquêtes combinées* plus substantielles dans lesquelles la partie «travail des enfants» a des incidences plus visibles sur la conception et les opérations de l'enquête de base (section 2.2.3). L'EMO constitue la base de la plupart des enquêtes combinées. Il en va de même pour plusieurs ETE modulaires, mais divers autres types d'enquêtes ont aussi servi de bases, comme l'indique le tableau 2.1.

Une autre possibilité, mais qui est rare, consiste à mener une *enquête liée,* dans laquelle l'entretien est séparé opérationnellement de celui de l'enquête de base, l'échantillon et certaines informations de fond étant cependant issus de l'enquête de base (section 2.2.3).

Résumé des différentes formes de liaison entre l'enquête de base et l'ETE

Modulaire (8)	Azerbaïdjan (2006), Cambodge (2006), Honduras (2002), Népal (1996), Nicaragua (2000), Ouganda (2000-2001), République Unie de Tanzanie (2000-01), Zambie (2000).
Combinée (5)	Costa Rica (2003), Kenya (1998-99), Turquie (1994), Turquie (1999), Ukraine (1999).
Liée (1)	Zimbabwe (1999).
Indépendante (15)	Bangladesh (2002-2003), Belize (2001), Cambodge (2001), République dominicaine (2002), Éthiopie (2001), Ghana (2001), Mongolie (2002-2003), Namibie (1999), Nigeria (1999), Pakistan (1996), Panama (2000), Philippines (2001), Portugal (1998), Afrique du Sud (1999), Sri Lanka (1999).

Le tableau 2.1 indique aussi que l'ETE est le plus souvent une enquête unique en une seule phase, bien qu'elle ait comporté dans un cas sur six des phases multiples, souvent au nombre de quatre correspondant aux trimestres de l'année, probablement pour appréhender les variations saisonnières. On notera qu'à la différence de l'EMO, qui recourt généralement à la rotation des échantillons, les phases de l'ETE seront probablement basées sur des échantillons indépendants ou qui ne se recoupent pas, ce qui permet de cumuler les données de l'ensemble des phases avec plus d'efficience.

Dans la plupart des circonstances, une ETE multiphase n'est pas une option accessible, si bien que nous n'approfondirons pas cet aspect dans ce manuel.

Les autres informations présentées dans ce tableau concernent la taille de l'échantillon, sa division en un certain nombre de grappes et le prélèvement par grappe. La taille de l'échantillon varie largement, de 6.000 à 48.000 dans les cas étudiés. La seule exception est celle d'un court module d'ETE combiné avec une très vaste opération de listage des ménages (Pakistan 1996, Section 2.3).

Le prélèvement par grappe est encore plus variable, dans une fourchette de 5 à 50 (avec la seule et même exception mentionnée ci-dessus). Dans une large mesure, cette variation peut être simplement le résultat d'un manque d'information sur les effets du plan de sondage et les paramètres des coûts, ou du défaut d'analyse et d'utilisation de cette information dans la conception (sections 3.4.3, 7.7, 7.8). Dans certains rapports d'enquête, l'information sur le nombre de grappes et le prélèvement par grappe n'est même pas disponible.

Liaison entre l'ETE et l'EET

Le tableau 2.2 classe les types de liaison entre les composantes ETE et EET. Il indique également le champ fondamental de l'EET: pour l'essentiel, il s'agit de savoir si elle concerne en premier lieu le travail des enfants (autrement dit les enfants qui sont au travail) ou si elle porte plus généralement sur les enfants ou les activités des enfants.

La forme prédominante est de loin celle de l'opération ETE-EET *intégrée*. Cela veut dire que les informations sur l'ampleur du travail des enfants (la composante ETE) et des informations plus détaillées sur les enfants que l'on trouve au travail sont les unes et les autres recueillies au cours d'une «seule et même opération d'entretiens», au sens des explications fournies dans la section 2.4. Généralement (ou peut-être dans tous les cas qui se sont présentés jusqu'ici), cela n'a impliqué aucun sous-échantillonnage entre les deux composantes.

Sans aucun doute, il peut être plus commode et plus rapide de couvrir les deux composantes en une seule opération. Cela peut aussi être considéré comme plus économique. Malheureusement, il n'en va pas nécessairement ainsi. L'ETE et l'EET posent des conditions très différentes en matière de conception et de taille des échantillons, comme nous le verrons aux chapitres 4 et 5. Une conception intégrée sans sous-échantillonnage intermédiaire peut souvent impliquer qu'un échantillon est trop petit pour l'ETE et/ou inefficient et pourtant trop grand pour la composante EET. Un échantillon trop grand pour l'EET aura très probablement un impact négatif sur la qualité des données complexes que l'enquête est censée recueillir.

Une solution de substitution réside dans les enquêtes *liées* qui permettent un certain degré de séparation opérationnelle entre l'ETE et l'EET (section 2.5), bien que cette disposition n'ait jusqu'ici été utilisée que dans seulement six des 31 enquêtes présentées dans le tableau 2.2. Il existe une grande diversité de formes de liaison possibles (la section 2.5 en présente 13 possibilités). Dans la pratique, la plupart des cas mentionnés dans le tableau 2.2 ont impliqué la prise en compte pour l'EET de la totalité ou d'un sous-échantillon des ménages comportant un enfant qui travaille, tels que l'ETE a permis de les identifier.

En ce qui concerne le champ fondamental de l'EET (qu'elle soit intégrée ou liée à l'ETE), elle est le plus souvent concernée essentiellement par l'*activité économique* des enfants qui travaillent. Toutefois, presque la moitié d'entre elles ont un champ plus vaste. On trouve un certain nombre d'enquêtes sur les *activités des enfants*, qui couvrent tous les types d'activités des enfants, y compris les activités non économiques; viennent ensuite les *enquêtes sur les enfants,* encore plus vastes, qui couvrent en plus d'autres domaines, tels que la santé, le logement et les conditions de vie des enfants (section 2.6). Ce champ élargi impose de couvrir une population plus large: pour l'essentiel, la totalité des enfants. Par conséquent, dans ce cas, l'«échantillon EET» devrait ressembler à celui de l'ETE, et pourtant une différence substantielle de taille d'échantillon serait normalement souhaitable, à cause des différences d'objectifs et de complexité existant entre les deux composantes. Les procédures de sous-échantillonnage du type qui est examiné au chapitre 4 peuvent en ce sens être mieux appropriées aux EET dont le champ est le plus vaste. Celles qui sont examinées au chapitre 5 sont mieux adaptées à l'EET «conventionnelle» sur l'activité économique des enfants.

Tableau 2.1: Enquêtes sur le travail des enfants: de l'enquête de base à l'ETE - Exemples de diversité des structures

Enquête	Année	(1) Liée à l'ÉTÉ	(2) si	(3) taille de l'échantillon et des grappes		
			multi-phases	n=a*b	a	b
Azerbaijan	2006	Modulaire (EMO)		17 000	850	20
Portugal	1998	Indépendante		25 000	1 150	22
Turquie	1994	Combinée (EMO)		13 500		?
Turquie	1999	Combinée (EMO)		20 000		?
Ukraine	1999	Combinée (EMO)	4 phases	48 000		?
Bangladesh	2002-03	Indépendante		40 000	1 000	40
Cambodge	1996	Modulaire (enquête socio-economique)	2 phases	9 000	750	12
Cambodge	2001	Indépendante		12 000	6 000	20
Mongolie	2002-03	Indépendante (même échantillon que EMO)	4 phases	12 000	1 200	10
Nepal	1996	Modulaire (enquête migration et emploi)		20 000	600	33
Pakistan	1996	Indépendante *uniquement une enquête sur listage pour identifier les ménages cibles*		140 000	1 860	75
Philippines	2001	Indépendante		27 000	2 250	12
Sri Lanka	1999	Indépendante	4 phases	15 000	1 000	15
Ethiopie	2001	Indépendante		44 000	1 250	35
Ghana	2001	Indépendante		10 000	500	20
Kenya	1998-99	Combinée: EMO, enquête sur le secteur informel, enquête sur le travail des enfants		13 000	1 100	12
Namibie	1999	Indépendante *uniquement une enquête sur listage pour identifier les ménages cibles*		8 000	270	30
Nigeria	1999	Indépendante		22 000	2 200	10
Afrique du Sud	1999	Indépendante		26 000	900	30
Tanzanie	2000-01	Modulaire (EMO)	4 phases	11 000	220	50
Ouganda	2000-01	Modulaire (enquête démographique et sanitaire)		8 000	300	27
Zambie	2000	Modulaire (enquête de multiples indicateurs)		8 000	360	22
Zimbabwe	1999	Liée (indicateur moniteur (IM)-EMO) *IM-EMO fournit les listes d'enfants dans tous l'échantillon de ménages*		14 000	400	35
Belize	2001	Indépendante		6 000	200	30
Costa Rica	2003	Combinée (enquête auprès des ménages polyvalente)		11 000	?	?
Rep. Dominicaine	2000	Indépendante		8 000	800	10
Honduras	2002	Modulaire (enquête permanente polyvalente)		9 000	1 800	5
Nicaragua	2000	Modulaire (ad-hoc EMO)		8 500	1 700	5
Panama	2000	Indépendante		15 000	1 500	10
Georgie, Romanie: similaire à l'Ukraine						

Un point d'interrogation indique que l'auteur ne disposait pas de l'information issue du rapport d'enquête ou d'une autre documentation.
Source: A partir des rapports nationaux des enquêtes sur le travail des enfants.

Tableau 2.2: Enquêtes sur le travail des enfants: de l'ETE à l'EET - Exemples de diversité des structures

Enquête	Année	(1) Basée sur l'ETE	(2) ETE à EET	(3) sous échantillonnage	(4) portée de l'EET
Azerbaijan	2006	Modulaire	Liée	sous échantillonnage des zones et des ménages ayant des enfants travailleurs n=4 000 a=400 b= 10 + plan spécifique pour les enfants réfugiés	Travail des enfants (TE)
Portugal	1998	Indépendante	Intégrée		Enquête sur les enfants
Turquie	1994	Combinée	Intégrée		TE
Turquie	1999	Combinée	Intégrée		TE
Ukraine	1999	Combinée	Intégrée		Enquête sur les enfants
Bangladesh	2002-03	Indépendante	Intégrée		TE +questionnaire des employeurs
Cambodge	1996	Modulaire	Intégrée		Enquête sur les enfants +questionnaire des employeurs
Cambodge	2001	Indépendante	Intégrée		Enquête sur les enfants
Mongolie	2002-03	Indépendante	Intégrée		Enquête sur les activités des enfants
Nepal	1996	Modulaire	Intégrée		TE
Pakistan	1996	Indépendante	Liée	tous les ménages comportant des enfants travailleurs (pas de sous échantillonnage) n=10 500 a=1 400 b=8	TE
Philippines	2001	Indépendante	Liée	tous les ménages comportant des enfants travailleurs (pas de sous échantillonnage)	TE
Sri Lanka	1999	Indépendante	Intégrée		child activity Enquête
Ethiopie	2001	Indépendante	Intégrée		TE +éducation
Ghana	2001	Indépendante	Intégrée		Enquête sur les activités des enfants
Kenya	1998-99	Modulaire	Intégrée		TE
Namibie	1999	Indépendante	Liée	tous les ménages comportant des enfants travailleurs (pas de sous échantillonnage)	TE
Nigeria	1999	Indépendante	Intégrée		Enquête sur les activités des enfants + Enquête sur les enfants de la rue
Afrique du Sud	1999	Indépendante	Liée	sous échantillonnage des ménages ayant des enfants travailleurs (toutes zones confondues)	TE
Tanzanie	2000-01	Modulaire	Intégrée		TE
Ouganda	2000-01	Modulaire	Intégrée		TE
Zambie	2000	Modulaire	Intégrée		TE

Enquête	Année	(1) Basée sur l'ETE	(2) ETE à EET	(3) sous échantillonnage	(4) portée de l'EET
Zimbabwe	1999	Liée (IM-EMO)	Intégrée		TE
Belize	2001	Indépendante	Intégrée		Enquête sur les enfants
Costa Rica	2003	Combinée	Intégrée		Enquête sur les activités des enfants
Rep. Dominicaine	2000	Indépendante	Intégrée		Enquête sur les activités des enfants
Honduras	2002	Modulaire	Liée	2 à 5 sous échantillonnage des ménages de toutes zones d'échantilonnage n=3 600 a=1 800 b=2	TE
Nicaragua	2000	Modulaire	Intégrée		TE
Panama	2000	Indépendante	Intégrée		Enquête sur les enfants
Georgie, Romanie: similaire à l'Ukraine					

Source: Compilation de Rapports d'enquêtes nationales sur le travail des enfants

2.2. Les relations de l'ETE avec l'enquête de base

Les structures des enquêtes sur le travail des enfants peuvent varier en fonction de leur contexte. Les enquêtes peuvent différer par les modalités et le degré de leur liaison avec des enquêtes portant sur d'autres thèmes.

Dans la plupart des pays, il n'est pas établi d'enquêtes régulières sur le travail des enfants: des opérations uniques ou occasionnelles ont seulement été entreprises. Ces enquêtes ont été organisées sous deux formes principales:

1. Les enquêtes modulaires
2. Les enquêtes indépendantes sur le travail des enfants

2.2.1. Les enquêtes modulaires

La collecte d'informations sur le travail des enfants en conjonction avec une enquête à large base sur la population générale telle que l'enquête sur la main-d'œuvre (EMO) peut généralement prendre la forme d'un module de questions sur le travail des enfants lié à l'EMO.

Les informations essentielles de l'ETE, à savoir les estimations de la proportion des enfants de diverses catégories qui sont engagés dans le travail des enfants, peuvent être obtenues en abaissant l'âge limite inférieur des questions de l'EMO sur l'activité économique. Cette possibilité existe quand le travail des enfants est défini selon le concept standard (EMO) de l'activité économique. Des questions différentes et généralement plus élaborées seront nécessaires avec une interprétation différente de ce qu'il faut entendre par «travail des enfants».

L'attrait majeur des enquêtes modulaires sur le travail des enfants est qu'elles offrent un arrangement économique et commode pour l'obtention d'informations essentielles sur

le travail des enfants. Les enquêtes modulaires sur le travail des enfants permettent également d'utiliser les questions énumérées dans l'enquête de base comme des variables explicatives et classificatrices dans l'analyse des données sur le travail des enfants.

Mais, comme l'indique le *Manuel du SIMPOC*, les enquêtes modulaires présentent deux problèmes potentiels:

- Le nombre et la précision des questions relatives au travail des enfants qui peuvent raisonnablement s'insérer dans des opérations essentiellement concernées par d'autres thèmes sont très limités.
- Pour assurer une haute qualité des données, les différents sujets d'enquête doivent être compatibles en termes de concepts, de définitions, de méthodes d'enquête, de périodes de référence, de couverture et de conditions de conception. Cette nécessité de compatibilité peut demander des compromis qui limiteront par la suite l'utilité des données obtenues sur le travail des enfants.

2.2.2. Les enquêtes indépendantes sur le travail des enfants

Les enquêtes indépendantes portent exclusivement ou essentiellement sur les sujets du travail des enfants. Une enquête indépendante est donc caractérisée par sa focalisation sur un sujet unique, et par un degré considérable de séparation des autres enquêtes dans sa conception et son exécution. Cette séparation permet souvent un meilleur contrôle et une plus grande flexibilité dans la conception et les opérations de l'enquête sur le travail des enfants. Le contenu, la conception et les procédures de l'enquête peuvent être choisis pour répondre plus précisément à ses besoins fondamentaux.

Si les points forts des enquêtes indépendantes sont une focalisation plus précise sur le thème du travail des enfants et une flexibilité dans la conception et les opérations, l'inconvénient majeur réside dans leur coût élevé. Il est difficile d'entretenir des enquêtes indépendantes afin qu'elles constituent une source d'informations régulières sur le travail des enfants.

Une enquête indépendante suit généralement une opération de listage à partir de laquelle on sélectionne l'échantillon de ménages ou d'adresses de l'enquête. Toutefois, la focalisation des enquêtes indépendantes sur un sujet unique n'exclut pas la coordination opérationnelle et le partage des moyens matériels et des dispositifs avec d'autres enquêtes, ou l'utilisation d'une couverture, de concepts, de définitions et de classifications. Dans la pratique, le statut d'enquête indépendante est en fait une affaire de degré. Une enquête par sondage n'est pratiquement jamais conçue et exécutée dans un isolement complet par rapport à d'autres enquêtes contemporaines.

2.2.3. Autres dispositifs: Les enquêtes combinées et liées

Dans le tableau 2.1, la plupart des enquêtes sur le travail des enfants ont été classées comme étant indépendantes ou modulaires. Dans quelques cas, nous avons identifié des sous-types qui spécifient plus précisément la structure particulière de l'enquête. Ce sont les enquêtes «combinées» et les «enquêtes liées».

Les enquêtes combinées

Celles-ci constituent une version plus complète de l'enquête modulaire. Quand on indique qu'un ensemble de questions sur le travail des enfants a été attaché en tant que «module» à une enquête de base existante, on comprend généralement que le nombre de ces questions additionnelles est assez faible pour n'avoir pas d'incidence significative sur la collecte de données pour l'enquête de base. Dans tous les cas, la conception de l'échantillon est déjà fixée, ou déterminée, pour refléter presque exclusivement les besoins de l'enquête de base.

Nous utilisons le terme «enquête combinée» pour indiquer la situation dans laquelle les questions sur le travail des enfants constituent un ajout substantiel à l'enquête existante au point d'influer sur les opérations de collecte de données, et/ou sur la conception de l'enquête et la taille de l'échantillon sont également influencées par les besoins de la composante «travail des enfants».

Il en a été ainsi, par exemple, des deux enquêtes sur le travail des enfants menées en Turquie et des enquêtes effectuées en Ukraine, au Kenya et au Costa Rica. Les encadrés 2.2-2.3 donnent quelques précisions sur deux de ces enquêtes: Costa Rica 2003 et Turquie 1994.

Dans l'enquête «combinée» menée au Costa Rica, la composante travail des enfants a pu influencer la conception globale de l'échantillon: «Compte tenu de l'intérêt porté par le pays à l'obtention de données sur le travail des enfants grâce à des modules spéciaux, l'échantillon national a été distribué par secteur (urbain et rural) selon les variations du taux de participation économique de la population âgée de 5 à 17 ans. Ce taux a été estimé sur la base du module d'enquête sur le travail des enfants de 1995.» (Rapport d'enquête, Costa Rica 2003.)

Dans le cas de l'enquête combinée (EMO+ETE) menée en Turquie, la conception de l'échantillon a été entièrement préétablie pour l'enquête auprès des ménages sur la main-d'œuvre. L'enquête sur le travail des enfants a impliqué un «sous-échantillonnage», qui a été réalisé simplement en appliquant la composante travail des enfants pendant seulement une des deux phases d'EMO de l'année. Néanmoins, le nombre substantiel des questions additionnelles sur le travail des enfants qui sont intervenues signifiait que l'enquête résultante pouvait plutôt être décrite comme une «opération combinée (EMMO+ETE)». Comme l'indique l'encadré 2.3: «L'application simultanée de l'EMMO et de l'ETE signifiait que les enquêtés avaient à répondre à un grand nombre de questions. Ces interminables questionnaires ont souvent ennuyé les enquêtés, qui sont devenus assez réticents à répondre aux questions vers la fin de l'enquête.» (Rapport d'enquête, Turquie 1994.)

Les enquêtes liées

Une ETE liée est une enquête dépendante de l'enquête de base (comme l'EMO) pour son échantillon et éventuellement pour d'autres informations, mais par ailleurs séparée de cette dernière sur le plan opérationnel dans le cadre de la collecte de données.

Une telle variante est représentée, par exemple, par l'enquête menée au Zimbabwe. L'enquête de base est une enquête de suivi d'indicateurs sur la main-d'œuvre. Exceptionnellement, cette enquête ne fournit pas seulement des listes d'adresses et de noms de chefs de famille dans les ménages de l'échantillon, mais aussi *les noms des enfants de 5 à 17 ans dans chaque ménage* pour l'enquête ultérieure sur le travail des enfants. L'ETE n'est donc pas une enquête indépendante. Elle est toutefois séparée de l'EMO sur le plan opérationnel et n'est donc pas un module attaché à cette enquête selon l'usage habituel. Nous avons appelé cet arrangement «enquête liée» pour le distinguer des trois arrangements décrits ci-dessus. Voir l'encadré 2.4, extrait du rapport d'enquête de 1999 pour le Zimbabwe: cet exemple illustre la grande diversité des arrangements qui sont possibles pour une enquête sur le travail des enfants.

Une enquête liée permet de quantifier le travail des enfants de manière plus détaillée et plus précise qu'il n'est normalement possible de le faire dans une enquête modulaire. Mais l'opération séparée qu'elle implique a évidemment un coût supplémentaire.

Autres commentaires

Quelques remarques complémentaires sur les illustrations mentionnées ci-dessus pourront être utiles. L'exemple du Costa Rica est très exceptionnel en ce sens que les modules d'enquête sur le travail des enfants doivent généralement accepter la taille d'échantillon et la conception de l'enquête de base existante à laquelle ils sont attachés. Les besoins d'échantillonnage liés aux questions sur le travail des enfants doivent être satisfaits en premier lieu en faisant varier les procédures de sous échantillonnage et les taux appliqués à l'échantillon existant de l'enquête de base.

Dans l'illustration concernant la Turquie, l'ETE n'est appliquée que pendant une seule des deux phases annuelles, ce qui fournit automatiquement un sous-échantillonnage à 50 pour cent des entretiens d'EMO à suivre dans le cadre de l'ETE.

Le point important déjà noté est que si les enquêtes nationales sur la main-d'œuvre sont en général relativement vastes, instituées et continues, les enquêtes sur le travail des enfants tendent à être des opérations plus limitées, occasionnelles ou uniques. Il est de plus en plus fréquent que les enquêtes sur la main-d'œuvre soient effectuées plus d'une fois par an: deux fois dans le cas de la Turquie mentionné ci-dessus, voire même de manière trimestrielle ou mensuelle dans certains des pays plus développés. Cette fréquence accrue est motivée à la fois par la nécessité de disposer de davantage de données actualisées et de suivre les variations à court terme et saisonnières de la population active.

Il est évidemment important de capter les variations saisonnières en relation avec l'activité économique des adultes (EMO) comme avec le travail des enfants (ETE). Pourtant, les données en rapport avec ces deux thèmes n'ont pas le même degré d'importance, et elles ont leur propre coût et autres contraintes pratiques. Même les données de base sur le travail des enfants sont rares mais très nécessaires, et, relativement parlant, moins de ressources sont disponibles pour améliorer cette situation. Par conséquent, il est souvent plus sage de donner plus de poids à l'obtention d'une position fiable sur la situation de base du travail des enfants dans le pays (y compris ses variations géographiques et sectorielles), que de consacrer, au stade actuel, des ressources accrues au suivi de ses variations saisonnières.

Il peut y avoir une énorme différence dans la somme des soins et du travail requis, et par conséquent dans le coût et le temps investis, pour la liaison d'une ETE à une enquête de base, en fonction du type d'information qui sera transféré entre les deux enquêtes au niveau micro. Disposer d'un échantillon commun d'adresses ou de ménages (l'ETE utilisant le même échantillon ou un sous-échantillon de l'EMO) exige seulement d'identifier, de sous-échantillonner et de redénombrer les ménages de l'EMO. Le décalage admissible entre les deux est déterminé par la mesure dans laquelle les adresses ou les ménages sont des unités stables dans le temps (les adresses sont habituellement plus stables que les ménages). L'option suivante consiste à utiliser l'EMO pour identifier les ménages comportant des enfants pour constituer la base ou l'échantillon de l'ETE. Ces unités sont un peu plus exposées au changement dans le temps (car elles sont également affectées par les changements qui interviennent dans la composition des ménages), et demandent davantage d'informations pour être transférées de l'EMO à l'ETE. Une procédure qui va beaucoup plus loin dans la même direction consiste à utiliser les noms et autres identifiants des enfants qui seront réellement suivis – mais en incluant tous les enfants, qu'ils travaillent ou non, dans le suivi. C'est l'option qui a été adoptée pour l'ETE du Zimbabwe 1999. Certaines enquêtes vont plus loin encore en suivant seulement les enfants qui ont été identifiés comme engagés dans le travail au cours de l'enquête antérieure de plus grande envergure.

Voir la section 2.5 pour un examen plus détaillé des différentes options de liaison entre les différentes composantes de l'enquête.

Encadré 2.2: Exemples de combinaisons EMO-ETE: Costa Rica 2003

La conception de l'échantillon correspond à une approche probabiliste, sectorielle, stratifiée et en deux degrés. …. Elle exige une autopondération dans chaque strate. Les secteurs de dénombrement ayant les mêmes probabilités, l'autopondération a été maintenue en choisissant une fraction fixe d'unités d'habitation durant la seconde phase – un quart (1/4) dans les secteurs de dénombrement urbains et un tiers (1/3) dans les secteurs ruraux. … La taille et la distribution de l'échantillon ont été calculées comme suit:

- La taille d'échantillon nécessaire pour obtenir des estimations de taux de chômage de 5 %, avec une marge d'erreur de 1 %, un niveau de confiance de 95 % et un effet du plan de sondage de 2,45, a été calculée.
- Quatre domaines d'enquête ayant été établis (Région centrale urbaine, Région centrale rurale, Autres urbains et Autres ruraux), la taille d'échantillon pour l'ensemble du pays a été obtenue en multipliant par quatre le résultat de l'opération mentionnée ci-dessus.
- Compte tenu de l'intérêt porté par le pays à l'obtention de données sur le travail des enfants grâce à des modules spéciaux, l'échantillon national a été distribué par secteur (urbain et rural) selon les variations du taux de participation économique de la population âgée de 5 à 17 ans. *Ce taux a été estimé sur la base du module d'enquête sur le travail des enfants de 1995* [souligné par nos soins].
- L'échantillon obtenu pour chaque secteur a été distribué par région, selon la variabilité relative du nombre des personnes sans emploi.
- La taille d'échantillon a été calculée en utilisant la procédure suivie pour les échantillons aléatoires simples. L'échantillon en question étant complexe, un ajustement a été effectué pour prendre en compte l'effet du plan de sondage. …

Source: National Report on the Results of Child and Adolescent Labour Survey in Costa Rica, 2003. Bureau International du Travail, 2004 (ISBN 92-2-115904-3).

Encadré 2.3: Exemples de combinaisons EMO-ETE: Turquie 1994

La phase d'octobre 1994 de l'Enquête auprès des ménages sur la main-d'œuvre (EMMO), qui est menée par l'Institut national de la statistique (INS) depuis 1988 sur une base semi-annuelle, incluait deux questionnaires additionnels destinés à évaluer pour la première fois l'incidence du travail des enfants en Turquie. Dans ce pays, la population active potentielle est définie comme étant constituée par les personnes âgées de plus de 12 ans ne présentant pas de handicap physique ou mental. … Cela a nécessité de mener une enquête spéciale sur le travail des enfants âgés de 6 à 14 ans. … Dans la phase d'octobre de l'EMMO 1994, on a appliqué aussi bien l'EMMO que l'ETE aux ménages comportant au moins un enfant âgé de 6 à 14 ans.

L'EMMO consiste en deux questionnaires: les formulaires A et B. La première partie du formulaire A a couvert les caractéristiques démographiques des membres du ménage, et la seconde partie la situation de l'emploi du chef du ménage. Les questions d'enquête de la seconde partie du formulaire A sont répétées dans le formulaire B, cette fois pour étudier la situation de l'emploi de tous les membres du ménage au-dessus de l'âge de 6 ans.

L'ETE, appliquée en tant que partie intégrante de l'EMMO d'octobre 1994, consiste également en deux questionnaires: les formulaires C et D.

Le formulaire C est utilisé pour déterminer les diverses caractéristiques du ménage, comme celles de son habitation (type d'habitation, surface habitable, etc.), l'historique migratoire des membres du ménage, les revenus mensuels du ménage, le nombre de membres du ménage employés, et la situation des enfants en matière d'emploi et d'éducation, grâce à des entretiens avec le chef du ménage - ou un autre membre du ménage en l'absence de celui-ci.

Le formulaire D est utilisé pour obtenir des informations plus détaillées concernant la situation des enfants en matière d'éducation et d'emploi par des entretiens directs avec les enfants eux-mêmes. Outre les questions habituelles concernant par exemple le nombre d'heures de travail par semaine, le type d'activité économique exercée, la situation relative à l'emploi et les gains mensuels, sont posées des questions attitudinales portant par exemple sur l'appréciation par les enfants de leurs conditions de travail, leurs relations avec l'employeur et leurs aspirations pour l'avenir. En plus des questions portant sur les activités économiques des enfants, d'autres visent à évaluer leurs activités domestiques (tâches ménagères) et leur temps de jeu.

Les problèmes suivants ont été rencontrés pendant l'enquête pilote:

(1) *L'application simultanée de l'EMMO et de l'ETE signifiait que les enquêtés avaient à répondre à un grand nombre de questions. Ces interminables questionnaires ont souvent ennuyé les enquêtés, qui sont devenus assez réticents à répondre aux questions vers la fin de l'enquête* [souligné par nos soins].

(2) Certains enquêtés ont semblé réticents à fournir des informations concernant leurs jeunes enfants travailleurs.

(3) Les questions sur la migration n'ont pas reçu de réponse claire …

(4) La communication avec les jeunes enfants a posé certains problèmes.

(5) Les ménages ont été très réticents à fournir des informations concernant leurs revenus.

(6) L'ETE étant appliquée dans le cadre de l'EMMO, il a fallu répéter certaines questions plusieurs fois pour assurer un enchaînement logique. Le fait d'avoir à répondre à nouveau à la même question a souvent ennuyé les enquêtés.

Source: Child Labour 1994. State Institute of Statistics Prime Ministry, Republic of Turkey, 1997 (ISBN 975-19-1635-6).

Encadré 2.4: Exemple de liaison EMO-ETE: Zimbabwe 1999

L'Enquête sur le travail des enfants (ETE) était un module de l'Enquête de Suivi d'Indicateurs sur la Main-dOeuvre (ESIMO) menée par le Bureau Central de la Statistique (CSO) en juin 1999. Cette dernière était une composante des enquêtes permanentes et intégrées de suivi d'indicateurs menées par le CSO auprès des ménages depuis 1993. Ces enquêtes ont un contenu en données déterminé et limité conçu pour assurer le suivi des conditions de vie. À cet égard, l'Enquête sur le travail des enfants était destinée à être une étude approfondie présentant un tableau national de la nature, de l'ampleur et des causes du travail des enfants au Zimbabwe. ….

La base de sondage par secteur utilisé pour l'ETE a été l'échantillon de référence du Zimbabwe 1992 (ZMS 92) élaboré par le CSO à la suite du recensement de population de 1992. Le ZMS 92 comportait 395 secteurs de dénombrement (SD) stratifiés par province et par affectation des terres. Le ZMS 92 a été révisé en 1997 pour l'Enquête démographique interrecensement (ICDS).

….. Un plan d'échantillonnage stratifié à deux degrés a été appliqué, les unités d'échantillonnage étant les SD pour le premier niveau et les ménages pour le second. Au total, 395 SD ont été sélectionnés selon le principe d'échantillonnage avec probabilité proportionnelle à la taille (PPT), la taille correspondant aux membres des ménages dénombrés dans le recensement de population de 1992. La sélection des SD a été une opération systématique, à un niveau, effectuée de manière indépendante pour chacune des 34 strates. Au sein de chacun de ces SD, une opération complète de listage et de cartographie des ménages a été effectuée, constituant la base de l'échantillonnage du second niveau. Les ménages résultants ont été sélectionnés par échantillonnage systématique aléatoire. La technique de l'échantillonnage autopondéré a assuré à tous les ménages de chaque province une égale probabilité de sélection. Elle a également simplifié l'analyse des données collectées grâce au calcul direct des pourcentages, moyennes et taux des paramètres de population à partir de l'échantillon, sans avoir nécessairement à pondérer ou à extrapoler les facteurs.

Au total, 13 591 ménages ont été choisis sur une liste de 55 176 ménages.

Au cours de l'ETE, *pour chaque SD sélectionné, le nom du chef du ménage et les noms des enfants de 5 à 17 ans dans chaque ménage ont été fournis à chaque enquêteur pour lui permettre d'identifier aisément les répondants qualifiés* …. [souligné par nos soins].

Une liste des SD sélectionnés a été remise à chaque bureau provincial. Outre la liste des SD, une liste contenant les noms du chef du ménage et des enfants âgés de 5 à 17 ans dans chaque ménage, ainsi que l'adresse du ménage à interroger, a aussi été remise à chaque enquêteur.

Pour les ménages répondants, un questionnaire semblable à celui de l'ESIMO a été utilisé pour recueillir des données sur l'activité des enfants. Les thèmes de cette ESIMO et de l'ETE étant semblables, les questions pertinentes qui figuraient déjà dans l'ESIMO n'ont pas été répétées. Le questionnaire de l'ETE s'est donc concentré exclusivement sur les variables relatives aux activités des enfants et à leurs conditions de travail.

… Le temps d'entretien approximatif nécessaire pour remplir le questionnaire d'ETE était de 20 à 30 minutes par répondant. … La collecte de données a commencé le 20 septembre 1999 est s'est achevée le 30 septembre 1999. Chaque enquêteur a couvert en moyenne quelque 70 ménages pendant la période de collecte de données.

Source: 1999 National Child Labour Survey, Country Report. Ministry of Public Service, Labour and Social Welfare, Zimbabwe.

2.3. L'implication pour l'échantillonnage de la définition du «travail des enfants» utilisée

La composante ETE collecte généralement une série d'informations sur les caractéristiques des ménages, sur les caractéristiques de base de tous les membres du ménage, et sur les caractéristiques et les activités des enfants. L'information «de définition» de cet ensemble est celle qui permet d'identifier *l'activité des enfants en relation avec le travail,* sur la base de laquelle on peut effectuer des estimations concernant l'ampleur du travail des enfants. Au cours de la composante EET, les parents ou les tuteurs des enfants sont ensuite soumis à des questions complémentaires, souvent très détaillées, sur la nature et les conséquences de leur activité en relation avec le travail. La composante ETE assure donc une sélection pour l'identification des cas échantillonnés pour la composante EET. Cette relation s'applique quelle que soit la forme sous laquelle les composantes ETE et EET sont liées l'une à l'autre sur le plan opérationnel: elle peut varier de l'intégration complète à une liaison distendue.

La pratique des enquêtes varie en ce qui concerne l'acception donnée ici à «l'activité en relation avec le travail»: elle peut inclure seulement l'activité économique rémunérée au sens strict, ou l'ensemble de l'activité économique, rémunérée ou non; elle peut aussi inclure les tâches domestiques «importantes» définies de manière appropriée, ou même toute autre activité, comme les soins donnés à d'autres enfants ou à d'autres personnes du ménage, qui peuvent potentiellement avoir un impact sur le développement et le bien-être de l'enfant. De même, la période de référence de l'activité peut varier, allant par exemple de la semaine en cours ou passée à une période prolongée telle que l'année passée. Voir l'encadré 2.5 pour une illustration du sens donné à des termes tels qu' «activité» et «travail» des enfants dans une enquête particulière (Portugal 1998), et l'encadré 2.6 pour les recommandations internationales sur le même sujet (BIT 2004).

Bien entendu, ces choix des concepts de base de l'enquête sont avant tout basés sur des considérations de fond, déterminées par les objectifs de l'enquête, et non sur des considérations de conception de l'échantillon. Cela veut dire que ce que l'on entend par «travail des enfants» est une question de fond et, à ce titre, n'est apparemment pas un problème d'échantillonnage. Toutefois, dans la pratique, *la définition du travail des enfants adoptée dans l'enquête peut avoir des implications importantes pour le choix de la structure de l'enquête et de la conception de l'échantillon*. Il y a au moins deux raisons à cela.

La complexité de la définition

Premièrement, le volume et la complexité de l'information qui doit être collectée pour identifier les «enfants travailleurs» parmi l'ensemble des enfants influent sur le type et la taille de l'enquête requise à cette fin. Par exemple, si le travail des enfants est défini comme une «activité économique» standard selon le concept de «population active» de l'OIT, il peut être capté par une simple extension de la limite d'âge inférieure pour les entretiens menés dans une enquête à grande échelle comme l'EMO. Parfois,

l'information sur l'activité économique des enfants a même été obtenue grâce à une très vaste opération de listage des ménages (ex.: Pakistan, 1996) – toutefois, nous ne recommandons pas cette pratique à cause du coût élevé et de la faible qualité de cette procédure, sauf à prendre des dispositions spécifiques pour élargir l'opération de listing afin d'y inclure un «mini-questionnaire» pour l'identification du travail des enfants (chapitre 6). Par contre, une définition très complexe du travail des enfants, en termes de caractéristiques particulières incluant toutes les activités non scolaires des enfants et excluant les loisirs, exigerait des interrogations beaucoup plus élaborées sur l'usage qu'ils ont fait de leur temps. De telles informations ne peuvent pas être aisément incorporées dans une enquête à grande échelle existante telle que l'EMO. En tout cas, la taille d'échantillon résultant de cette incorporation sera probablement d'une dimension déraisonnable pour l'information complexe qu'il faudra collecter, en prenant en compte le coût, la qualité des données et les contraintes pratiques.

Le caractère restrictif

La seconde implication, pour l'échantillonnage, de la définition particulière du travail des enfants adoptée est la suivante. Ce choix détermine la proportion d'enfants classés parmi ceux qui sont engagés dans le travail, et par suite la taille d'échantillon résultante de *la partie de l'enquête consacrée aux investigations sur les conditions et les conséquences du travail des enfants* – cette partie pour laquelle la population de référence est celle des enfants travailleurs, et non de tous les enfants.

Ainsi, par exemple, une très large définition du travail des enfants pourrait signifier qu'une proportion importante d'enfants est enregistrée comme étant engagés dans le travail. Cela peut avoir pour résultat la création d'un échantillon de travailleurs enfants trop important pour étudier dans le détail les conditions et les conséquences du travail des enfants, et nécessiter un sous-échantillonnage de la totalité des enfants identifiés comme «travailleurs» dans l'échantillon d'origine avant leur inclusion dans l'étude détaillée sur les enfants travailleurs.

Par contre, une définition trop restrictive ou trop étroite du travail des enfants peut signifier que seule une faible proportion des enfants est considérée comme engagés dans le travail. Un échantillon d'enfants travailleurs suffisamment important pour l'étude des conditions et des conséquences du travail des enfants peut imposer d'accroître le nombre total des enfants étudiés dans l'échantillon original (de l'ETE) en augmentant la taille de celui-ci. Cela peut ne pas être toujours possible, par exemple si l'échantillon original provient d'une enquête existante avec une taille d'échantillon prédéterminée. Il peut être nécessaire de trouver des solutions alternatives. En voici un exemple. Dans une enquête sur le travail des enfants menée en Azerbaïdjan en 2006, l'échantillon d'EMO existant de près de 9 000 ménages ne pouvait pas fournir l'échantillon requis de plus de 5 000 ménages comportant des enfants travailleurs correspondant à la définition de l' «activité économique» standard. On a pu répondre de manière plus adéquate aux objectifs de l'enquête sur le travail des enfants en élargissant la définition du travail des enfants pour y inclure l'engagement dans des «travaux ménagers importants», augmentant ainsi le nombre des enfants ainsi définis comme

enfants travailleurs dans l'échantillon d'ETE, qui ont tous été inclus dans l'EET qui a suivi[5].

En résumé, les choix fondamentaux relatifs à ce qui est inclus dans le travail des enfants peuvent avoir des conséquences importantes pour l'échantillon résultant, établi en vue d'une enquête détaillée sur les enfants travailleurs. La raison en est que la taille d'échantillon réalisable pour la composante EET dépend de la mesure dans laquelle la définition adoptée de l'«activité en relation avec le travail» est plus ou moins large ou restrictive. Les tailles d'échantillon relatives en termes de nombre des enfants inclus dans les composantes EET et ETE sont conditionnées par la proportion d'enfants engagés dans une activité en relation avec le travail: dans une situation donnée, cette proportion dépend du choix de ce qui est inclus dans l'«activité en relation avec le travail».

Il est instructif de noter à ce propos comment le concept de travail des enfants est examiné dans le *Manuel du SIMPOC* (voir encadré 2.6).

La fonction de sélection de l'ETE

L'idée fondamentale du système ETE-EET est que l'ETE apporte des informations sur la base desquelles peut être identifié un échantillon d'enfants travailleurs plus efficient et mieux ciblé en vue de l'EET.

Pour l'essentiel, l'ETE identifie des unités (secteurs, ménages ou même enfants considérés individuellement) qui comportent des enfants travailleurs. Les unités ainsi identifiées peuvent contenir des «faux positifs» (travail des enfants signalé alors qu'il n'existe pas) comme des «faux négatifs» (travail des enfants existant non signalé). D'un point de vue fondamental mais aussi pratique, les faux négatifs présentent un problème beaucoup plus sérieux que les faux positifs. L'EET est censée être un instrument plus sensible pour une identification précise de la situation de travail des enfants. Les faux positifs issus de l'ETE peuvent être identifiés plus facilement au cours de la phase d'EET. Toutefois, les faux négatifs (les cas de travail des enfants non repérés) constituent une petite partie de l'importante population d'enfants et tendent à être beaucoup plus difficiles et coûteux à identifier. Ils seront complètement ignorés si l'EET se confine aux seuls cas identifiés comme positifs dans l'ETE.

---

[5] Le rapport d'enquête donne les informations méthodologiques suivantes.
«L'ETE a recueilli des informations sur 16 482 enfants âgés de 5 à 17 ans provenant de 8 785 ménages à travers le pays. … L'EET a étudié 10 543 enfants âgés de 5 à 17 ans provenant de 5 355 ménages. Si le but essentiel de l'EET était de recueillir des informations complémentaires sur les enfants travailleurs, tous les enfants des ménages couverts par l'enquête ont été interrogés, ce qui a produit un autre ensemble de données (en complément de celui de l'ETE) incluant à la fois les enfants travailleurs et non travailleurs. … Du fait que les enfants non travailleurs de l'échantillon d'EET partagent les mêmes caractéristiques des ménages que les enfants travailleurs – et peuvent donc être considérés comme plus exposés au travail que les enfants non travailleurs en général –, il est probable que les enfants non travailleurs de l'échantillon d'ETE sont plus représentatifs de la population des enfants non travailleurs dans son ensemble. Pour cette raison, cette étude s'est appuyée essentiellement sur les résultats de l'ETE dans l'analyse des principales caractéristiques des enfants travailleurs et des déterminants du travail des enfants.., ainsi que des déterminants de la fréquentation scolaire .., alors que les résultats de l'EET ont été utilisés pour évaluer les effets possibles du travail sur la santé et la fréquentation scolaire. …»

Une manière pratique de réduire ce problème consiste à utiliser une définition plus inclusive, c'est-à-dire plus large, du travail des enfants dans l'ETE, afin de permettre une définition plus étroite pour l'EET. Cela aura pour effet de réduire le nombre des faux négatifs de l'ETE *en fonction de la définition de l'EET* (censée être la définition qui présente un réel intérêt). La proportion de faux positifs identifiée dans l'EET pourra être utilisée pour ajuster convenablement les estimations de l'ampleur du travail des enfants issues de l'ETE.

Encadré 2.5: Implications de la terminologie et des définitions clés pour l'échantillonnage: Portugal 1998

… Le terme «activité» est utilisé au lieu de «travail» plus d'une fois dans ce texte, ce dernier terme étant utilisé dans un sens précis. Ainsi le terme «activité» a un usage très général, qui peut évoquer le sens économique du mot (travail), voire même des activités ménagères et scolaires. Ainsi, l'expression «activité économique» sert à designer des situations d'emploi par un tiers ou de travail familial non rémunéré, telle qu'elle est utilisée dans le domaine de l'emploi et des statistiques du chômage. Compte tenu de l'improbabilité de cette situation dans le cas d'enfants, le travail indépendant n'est pas envisagé. Par conséquent, au sens du questionnaire, «enfants ayant une activité économique» correspond au travail des enfants. Avec cette définition, le concept juridique restrictif de travail des enfants se trouve élargi pour designer tout ce qui peut être considéré comme enfreignant la législation actuelle et qui inclut un travail qui pourrait (de par ses caractéristiques) faire l'objet d'une relation contractuelle.

À leur tour, les activités ménagères sont elles aussi visées par ce questionnaire, à cause de l'effet négatif qu'elles pourraient avoir sur le développement d'un enfant et parce qu'elles pourraient même avoir des effets semblables à ceux du travail des enfants. Toutefois, elles ne constituent pas une activité au sens économique du terme, et moins encore dans le domaine du droit du travail. Les activités ménagères devraient donc être considérées en fonction de leur difficulté, de leur durée ou en tant que facteur faisant obstacle à la scolarité et à ses avantages.

Dans le questionnaire, les activités ménagères sont considérées comme étant celles qui sont menées avec une certaine régularité et à titre d'aide à la famille, selon les déclarations des enfants et des chefs des ménages. Il ne s'agit pas d'activités soumises à des horaires, mais de tâches légères accomplies dans le cadre d'un processus d'éducation familiale, telles que nettoyer ses vêtements, faire son lit ou débarrasser la table.

La nécessité de quantifier les différentes situations d'activité signifiait que l'activité économique était considérée comme un élément déterminant dans la classification des enfants. Par conséquent, dès lors qu'un enfant a une activité économique, quelle que soit son importance, il est inclus dans cette catégorie, même si elle appartient aux tâches ménagères.

Les enfants «sans activité» sont ceux qui n'ont pas d'activité au sens économique du terme. Toutefois, dans l'analyse du questionnaire, on utilise aussi l'expression «activités scolaires». Cela inclut la participation aux cours, à l'étude et aux devoirs faits à la maison, mais sans valeur économique. La fréquentation scolaire est donc quantifiée. Les activités scolaires sont celles qui sont menées par les enfants scolarisés, qu'ils aient ou non d'autres activités.

Source: Child Labour in Portugal: Social Characterisation of School Age Children and Their Families, 1998. Portugal Ministry of Labour and Solidarity (MTS).

Encadré 2.6: Sur le concept de travail des enfants : Manuel du SIMPOC (BIT, 2004).

Qu'est-ce que le travail des enfants ? Il n'y a pas de définition universellement acceptée du concept de travail des enfants. Les positions divergent de façon très prononcée parmi les chercheurs quant aux types d'activités qui peuvent être qualifiés de travail des enfants ...

Le nœud du débat est le suivant: quelle est la meilleure approche de définition du «travail» en ce qui concerne les enfants, c'est-à-dire adopter le critère général de la frontière de production ou adhérer à la frontière de production du système de comptabilité nationale des Nations Unies ? ...

[À un extrême, certains] observateurs soutiennent que le travail des enfants devrait seulement prendre en compte toutes ces activités économiques qui empêchent un développement normal de l'enfant en adulte responsable. Cette perspective réserve le terme de «travail des enfants» à l'emploi dans des activités économiques fatigantes ou dangereuses pour les enfants en bas âge ainsi que les enfants dans les pires formes de travail (PFTE). ...

Certains chercheurs adoptent cependant un concept plus large de «travaux effectués par les enfants», qui, selon la limite de production, comprend les activités non économiques usuelles telles que les travaux ménagers. .... Les adeptes d'une définition plus large, qui prend en compte les activités non économiques et les travaux ménagers, soutiennent que l'adhésion à une définition restreinte du travail des enfants comme un sous-ensemble des activités économiques comporte le risque de données biaisées selon le genre. ...

[À l'autre extrême, certains observateurs adoptent] le point de vue selon lequel toute activité extrascolaire ne relevant pas du divertissement constitue un travail des enfants. Ainsi le travail des enfants comprendrait les petits travaux domestiques après l'école, ou même l'aide dans les travaux ménagers comme nettoyer la maison ou veiller sur les petits frères ou petites sœurs.

L'approche du BIT/SIMPOC ... définit le travail des enfants comme un sous-ensemble des activités économiques (c'est-à-dire le travail) accomplies par les enfants. Cela est basé sur la définition du travail de l'OIT telle qu'elle s'applique aux personnes en âge de travailler, qui est liée à la limite de production fixée par le système de comptabilité nationale des Nations Unies. .... En commençant avec un cadre large et en évoluant vers un concept plus restreint, la série de définitions statistiques du «travail des enfants» adoptées par les chercheurs étudiant les enfants accomplissant diverses activités non scolaires pourrait comprendre:

- «les enfants dans les activités non économiques (y compris les travaux domestiques)» et les «enfants travailleurs»;
- «les enfants travailleurs» uniquement (c'est-à-dire accomplissant des activités économiques); et
- «le travail des enfants à abolir» tel qu'adopté dans l'approche du BIT/SIMPOC, à savoir un sous-ensemble des «enfants travailleurs». ...

Inclure les activités non économiques dans les estimations du travail des enfants, même en utilisant pour cela une définition standard internationale de ces activités, peut également se révéler problématique. [Un certain nombre de démarches sont nécessaires dans la mise en œuvre de cette entreprise:] élaborer des critères pour prendre en compte les tâches ménagères, définir des seuils pour les «travaux légers» [au-dessus desquels seulement une activité pourra être considérée comme un «travail»], élaborer des définitions des «travaux dangereux», établir des limites d'âge plus universellement acceptées pour les enfants enquêtés, et améliorer la conception du questionnaire.

…

Les estimations globales du travail des enfants à abolir du BIT

… En l'absence d'une définition universelle du travail des enfants, les estimations globales du travail des enfants à abolir du BIT menées en 2002 ont été guidées par des âges de base discutés dans les clauses des résolutions de l'OIT et les actes internationaux significatifs dont plusieurs pays sont signataires.

En supposant un âge minimum de 12 ans pour les travaux légers et un âge minimum de 15 ans pour l'admission à un emploi, le BIT a estimé l'incidence globale du travail des enfants à abolir en utilisant une mesure qui prend en compte les enfants suivants:

- les enfants de 5 à 11 ans accomplissant des activités économiques;
- tous les enfants travailleurs âgés de 12 à 14 ans, sauf ceux qui accomplissent des travaux légers; et
- tous les enfants âgés de 15 à 17 ans astreints à des travaux dangereux et aux PFTE [pires formes de travail des enfants] intrinséques.

Les travaux légers ont été définis comme étant des travaux non dangereux accomplis par des garçons et des filles de 12 ans au moins pendant moins de 14 heures par semaine (en moyenne deux heures par jour). Les travaux dangereux, d'autre part, comprennent les travaux accomplis pendant 43 heures et plus par semaine ainsi que les travaux de mine, de construction, et des professions spécifiques souvent considérées comme dangereuses dans plusieurs pays.

… Il faut noter que ces procédures de classification des enfants économiquement actifs sont spécifiques à cet exercice particulier d'estimation globale et ne représentent pas une recommandation de l'OIT pour des enquêtes statistiques nationales.

Source: Manuel du SIMPOC (extraits des pages 17, 21, 22, 27-31).

2.4. L'EET intégrée

Dans le tableau 2.2 (colonne 2) ont été identifiées deux formes de relation entre l'ETE et l'EET: une enquête intégrée et des enquêtes liées.

La distinction, évoquée dans la section 2.1, entre les deux types d'enquête, l'ETE et l'EET, n'implique aucunement qu'elles doivent (ou même puissent) être organisées comme deux *opérations* distinctes. En fait, dans la majorité des enquêtes nationales menées jusqu'ici, elles ont été complètement intégrées en une seule opération d'enquête – la même enquête couvrant les deux types d'objectifs différents.

2.4.1. Les caractéristiques de la conception intégrée

Par «EET intégrée», nous entendons que l'information nécessaire aux composantes ETE et EET est collectée au cours d'une seule et unique opération d'entretiens. On notera qu' «une seule et unique opération d'entretiens» n'implique pas nécessairement que l'on utilise un seul questionnaire intégré, ni que l'ensemble de l'entretien ait lieu durant une seule visite rendue au ménage, ni même que toute l'information soit obtenue d'un seul répondant dans le ménage. Les questionnaires multiples, les visites répétées auprès des ménages et la diversité des répondants (chef du ménage, parents ou tuteurs, enfants eux-mêmes, et parfois même employeurs et enseignants, etc.) peuvent en fait intervenir. «Une seule et unique opération d'entretiens» est censé indiquer que l'information nécessaire aux composantes ETE et EET est collectée en même temps ou du moins à intervalle rapproché et que toute l'information recueillie pendant la partie ETE de l'opération est directement disponible pour la composante EET (et n'est généralement pas répétée au cours de celle-ci).

Normalement, l'enquête intégrée prend la forme de questions d'EET attachées au questionnaire d'enquête et qui deviennent applicables si l'on constate que l'enfant concerné est engagé dans une activité en relation avec le travail au cours d'une période de référence spécifiée.

L'exécution simultanée des composantes ETE et EET (ou du moins l'absence de tout laps de temps notable entre celles-ci) implique que l'on ne peut normalement introduire aucun sous-échantillonnage de l'une à l'autre (autrement dit, que tous les enfants identifiés comme engagés dans le travail pendant la partie ETE sont soumis à la partie EET de l'entretien). En tout cas, il est souhaitable dans la pratique que tout sous-échantillonnage introduit dans l'enquête reste simple, comme l'application de la partie EET de l'entretien à un sous-ensemble présélectionné de phases de l'ETE ou de secteurs de dénombrement.

Dans la mesure où le sous-échantillonnage ne peut pas être (ou n'est pas) introduit entre les composantes EET et ETE d'une enquête intégrée, les tailles d'échantillon relatives en termes de nombre d'enfants des deux composantes sont déterminées entièrement par la proportion d'enfants engagés dans des activités en relation avec le travail; dans une situation donnée, cette proportion dépend du choix de ce qui est inclus dans les «activités en relation avec le travail», comme on l'a vu ci-dessus.

Dans une enquête intégrée, il s'agit de remplir à la fois les conditions de taille d'échantillon des composantes ETE et EET. Par suite, *même dans une enquête unique «intégrée», la distinction entre les deux types de composantes de l'enquête demeure conceptuellement utile: elle nous rappelle qu'une conception intégrée doit être un compromis entre deux types d'objectifs*. En termes pratiques, cela revient à exiger que l'ETE soit assez vaste pour fournir des estimations de l'ampleur du travail des enfants avec la précision requise, et aussi pour produire des nombres suffisants d'enfants travailleurs pour l'EET.

Malheureusement, dans la pratique, il n'est pas rare que la conception de l'enquête ait été déterminée de manière plus ou moins unilatérale, privilégiant un type d'objectif aux dépens d'un autre. Ainsi, certaines enquêtes ont été trop restreintes pour produire des estimations utiles de l'ampleur et de la distribution de la pratique du travail des enfants (la taille d'échantillon de la composante ETE était trop limitée), alors que d'autres ont eu une taille trop importante pour permettre des recherches suffisamment approfondies sur les caractéristiques et les conséquences du travail des enfants (la taille d'échantillon de la composante EET était trop importante d'un point de vue pratique). Par contre, il existe aussi des exemples dans lesquels la taille de l'échantillon, tout en étant adéquate pour l'ETE, s'est révélée inadéquate pour fournir à la composante EET un nombre de cas suffisant pour des recherches détaillées sur les activités de travail des enfants.

Il vaut la peine de réaffirmer qu'avec une conception intégrée, l'échantillon de la composante EET est entièrement déterminé par les résultats de la sélection assurée par la composante ETE. Non seulement la taille de l'échantillon, mais aussi – ce qui est plus important – la qualité de la couverture de la population des enfants travailleurs dans la composante EET, sont déterminées par la qualité des questions de sélection dans la partie ETE pour l'identification de ces enfants.

Généralement, les considérations qui précèdent s'appliquent également aux enquêtes ETE et EET séparées mais «liées», qui seront évoquées dans la section suivante – bien que, dans ce cas, la séparation opérationnelle entre les deux enquêtes puisse permettre un certain contrôle de la couverture de l'EET et réduire sa dépendance par rapport à la qualité de la sélection opérée au cours de l'ETE. De plus, cette séparation facilite l'introduction d'un sous-échantillonnage entre les deux opérations.

2.4.2. Avantages et inconvénients d'une conception intégrée

Il peut y avoir des avantages pratiques et économiques à intégrer les composantes ETE et EET en une seule opération. Il est clair qu'il est moins coûteux et plus commode de collecter en une seule fois toute l'information requise. Cela constitue un avantage majeur et peut expliquer pourquoi, à ce jour, une grande majorité des enquêtes sur le travail des enfants a choisi la disposition intégrée.

Toutefois, la conception intégrée (ETE+EET) présente aussi quelques inconvénients par rapport à l'organisation de deux opérations séparées, bien que liées à des degrés divers.

- Une exécution intégrée tend à limiter l'information qui peut être collectée au cours de la composante EET, sans compromettre la qualité de la composante ETE résultant d'une charge de répondants accrue.

- Mis à part l'évidente limite supérieure imposée par le nombre de cas admissibles (enfants travailleurs) identifiés dans la composante ETE, il n'y a pas de contrôle indépendant de la taille et de la distribution de l'échantillon pour lequel est recueillie l'information de l'EET. Dans la pratique, l'introduction d'un sous-échantillonnage de l'ETE vers l'EET dans une conception intégrée n'est pas chose facile. Comme on l'a indiqué ci-dessus, elle est plus facilement réalisable quand les deux composantes sont séparées dans le temps.
- Le niveau du travail des enfants peut être distribué de manière très inégale sur le secteur de l'échantillon, ce qui introduit des variations considérables dans la charge de travail liée aux entretiens. Cela devient d'autant plus gênant lorsque la partie EET comporte de très longs entretiens avec des adultes ainsi qu'avec des enfants sur un plan individuel.
- La charge accrue associée à la partie EET peut avoir des conséquences sérieuses pour la qualité de la couverture de l'échantillon dénombré dans la partie ETE. En effet, pour réduire la charge de travail (souvent substantielle) impliquée par la partie EET de l'entretien, il arrive que l'on ait parfois tendance à sous-estimer les enfants travailleurs. En fait, cette tendance a été largement rapportée dans d'autres enquêtes comportant des modules assez lourds qui s'appliquent seulement à une partie de la population identifiée comme concernée au cours du stade précédent de l'entretien[6].

2.5. L'EET liée à l'ETE

Une solution de substitution à la conception intégrée consiste à conduire les composantes ETE et EET comme des opérations séparées. Toutefois, il ne peut s'agir d'enquêtes indépendantes (entièrement séparées), et elles doivent d'une certaine manière être liées l'une à l'autre. En effet, l'EET dépend de l'ETE pour l'identification de son échantillon d'enfants travailleurs, en termes d'identification des secteurs de l'échantillon d'ETE qui contiennent de fortes concentrations d'enfants travailleurs, comme des ménages qui comportent ces enfants ou des enfants travailleurs eux-mêmes.

L'échantillon d'EET peut donc être identifié de différentes manières sur la base des résultats de l'ETE, ce qui donne lieu à différentes formes de liaison entre les deux enquêtes.

2.5.1. Les différentes formes de liaison ETE-EET

L'EET peut être liée de plusieurs manières à une ETE effectuée antérieurement. La meilleure solution dépend de la situation et des objectifs spécifiques, mais essentiellement de deux facteurs: (1) *le laps de temps qui sépare les deux enquêtes;* (2) la *qualité de la sélection* assurée par l'ETE dans l'identification de la présence d'enfants travailleurs.

[6] Les enquêtes démographiques et sanitaires (EDS), par exemple, ont souvent décelé une tendance à sous-estimer les femmes en âge de procréer, ainsi que les nouveau-nés, quand l'entretien impliquait de recueillir des informations complémentaires pour ces populations.

La diversité des options de liaison ETE-EET se présente comme suit.

A. Concernant la sélection de l'échantillon parmi les secteurs de sondage de l'ETE, nous pouvons retenir la totalité ou choisir un échantillon:

1. des enfants travailleurs précédemment identifiés, considérés individuellement;
2. des ménages précédemment identifiés comme comportant un enfant travailleur;
3. des ménages précédemment identifiés comme comportant un enfant du groupe d'âge concerné;
4. de tous les ménages interrogés dans l'échantillon précédent;
5. ou éventuellement de tous les ménages sélectionnés dans l'échantillon précédent (y compris les non-répondants);
6. de tous les ménages listés dans les secteurs d'échantillon de l'ETE (autrement dit, obtenir un nouvel échantillon d'EET à partir des listes des ménages utilisées pour l'ETE); ou
7. des ménages à partir d'un relistage des secteurs d'échantillon de l'ETE (autrement dit, actualiser les listes par secteur avant de sélectionner un nouvel échantillon).

B. Concernant l' «admissibilité à l'inclusion» des secteurs d'échantillon de l'ETE, nous pouvons considérer comme «admissibles»:

8. les secteurs d'échantillon de l'ETE comportant au moins un enfant travailleur (ou, de manière équivalente, au moins un ménage comportant un enfant travailleur); la limite inférieure peut être élevée à «au moins x», où $x>1$;
9. les secteurs d'échantillon de l'ETE comportant au moins un enfant (ou au moins x enfants, $x>1$) dans le groupe d'âge concerné; ou
10. tous les secteurs d'échantillon de l'ETE.

C. Concernant la sélection des secteurs, nous pouvons:

11. prendre un sous-échantillon des secteurs d'échantillon de l'ETE;
12. prendre tous les secteurs d'échantillon de l'ETE; ou
13. choisir des secteurs additionnels pour l'EET, liés aux secteurs d'échantillon de l'ETE.

Certaines des options des trois ensembles A-C peuvent être combinées, par exemple:

- les ménages identifiés antérieurement comme comportant un enfant du groupe d'âge concerné (option 3),

- mais seulement dans un sous-échantillon de secteurs (option 11),
- sélectionnés dans des secteurs d'échantillon de l'ETE comportant au moins un enfant travailleur (option 8).

L'échantillonnage au sein des secteurs de l'ETE

L'option 1 assure le lien le plus étroit entre les deux enquêtes: l'échantillon d'EET est limité aux enfants individuellement identifiés comme engagés dans le travail des enfants au cours de l'ETE. Cela rend cette option presque semblable à une conception ETE-EET intégrée, excepté la possibilité d'un sous-échantillonnage entre les deux opérations à cause de leur séparation opérationnelle dans le temps.

On notera que parmi les options 1 à 7, toutes à l'exception de 1 impliquent une nouvelle identification des enfants travailleurs dans les ménages inclus dans l'échantillon d'EET. Avec l'option 1, l'échantillon d'EET est limité à l'ensemble particulier des enfants travailleurs identifiés au cours de l'entretien d'ETE.

Avec cette option, tout enfant travailleur non identifié pendant l'ETE reste dans la même situation pendant l'opération d'EET ultérieure. L'erreur opposée – généralement beaucoup moins importante – qui consiste à identifier comme travailleurs des enfants non travailleurs pourra évidemment être décelée au cours de l'opération ultérieure.

Avec les options 2 à 7, l'EET possède un plus grand potentiel pour affiner les estimations du travail des enfants issues de l'ETE.

Les catégories allant de 13 à 1 représentent une liaison de plus en plus étroite entre les deux enquêtes. En remontant de 13 à 1, et notamment de 7 à 1, les options deviennent généralement:

- plus restrictives (l'échantillon d'EET est de plus en plus restreint aux unités identifiées dans l'échantillon d'ETE);
- mieux centrées (sur la population particulière concernée, à savoir les enfants travailleurs); et
- plus efficientes pour l'exploitation de l'information déjà collectée ou la conception déjà mise en œuvre dans l'ETE (donc plus rentables et moins pesantes pour les enquêtés).

Mais ces options exigent:

- de collecter davantage d'informations au cours de l'ETE, puis de les introduire dans l'EET (l'opération devient donc plus coûteuse et demande davantage de temps); elles sont aussi
- plus sensibles aux changements dans le temps (donc moins appropriées en présence de laps de temps prolongés entre les deux enquêtes).

L'option 1 ne couvre pas les enfants de la population (et non pas seulement de l'échantillon d'ETE) qui ont commencé à travailler après l'époque du dénombrement de l'ETE. Cela introduit une distorsion, à moins que l'intervalle de temps entre les deux enquêtes ne soit très court, de l'ordre de quelques semaines au plus. Par contre, l'option 7

exige seulement la stabilité des secteurs de dénombrement, et peut donc être utilisée même quand les deux enquêtes sont largement séparées dans le temps.

Les options 1 à 3 exigent des informations sur certaines caractéristiques des ménages ou des personnes; les options 4 à 7 exigent seulement des informations sur les adresses des ménages.

On notera que les options 1 et 2 sont très différentes en ce qui concerne l'information nécessaire à leur application. Les ménages ou les adresses spécifiés, comme dans l'option 2, sont beaucoup plus faciles à identifier (et à identifier correctement) que les enfants spécifiés, comme dans l'option 1.

Les ménages sont probablement beaucoup plus stables sur le point de savoir s'ils comportent des enfants, que sur celui de savoir s'ils comportent ou non des enfants *travailleurs*. Cela veut dire que l'option 3 est plus appropriée quand l'intervalle entre les enquêtes n'est pas très court, que l'une ou l'autre des options 1 ou 2. De plus, l'option 3 évite les visites auprès des ménages qui sont probablement sans enfants et peut donc être sensiblement plus efficiente que l'option 4.

L'exclusion de certains secteurs d'ETE

La différence entre les options 9 et 10 peut être minime quand les secteurs de l'échantillon sont relativement vastes (dans la mesure où la plupart des secteurs comportent au moins un enfant). L'option 9 est couramment utilisée pour éviter des efforts inutiles sur des secteurs connus pour la forte improbabilité de la présence d'un enfant du groupe d'âge spécifié.

On prendra garde au risque de distorsion lié à l'option 8 – inclure seulement les secteurs comportant un ménage avec un enfant travailleur – si l'intervalle entre les deux enquêtes est important (de l'ordre de plusieurs mois).

Il peut toutefois y avoir de fortes raisons pratiques de choisir cette option. En général, le travail des enfants est très inégalement distribué sur le plan géographique. On en trouve souvent d'importantes concentrations dans une proportion de secteurs limitée, contre une faible prévalence dans une grande proportion de secteurs. La proportion des secteurs qui ne signalent aucun travail des enfants peut également être élevée. Ce modèle de distribution – et même, dans une large mesure, la position de secteurs particuliers à l'intérieur de celui-ci – est déterminé par les conditions socio-économiques prévalentes, qui tendent à persister dans le temps.

Par conséquent – en dépit de la mise en garde ci-dessus concernant le choix de l'option 8 –, l'information relative à la classification des secteurs d'ETE en fonction du niveau du travail des enfants peut être assez stable, donc très précieuse pour déterminer la conception de l'échantillon et la stratégie de mise en œuvre de l'EET[7].

Outre qu'elle sert à établir la structure générale du sous-échantillonnage de l'EET à partir des secteurs de l'ETE, l'information relative au niveau du travail des enfants rapporté dans

[7] Il est toutefois très important de noter que de nombreuses formes de travail des enfants sont sujettes à d'importantes variations saisonnières (période des vacances scolaires, saisons des récoltes, saisons touristiques, etc.). Si l'on utilise des informations d'une enquête à une autre, ces deux enquêtes devraient être menées pendant la même saison ou pendant des saisons semblables.

celle-ci est souvent utile pour déterminer le seuil en dessous duquel les secteurs peuvent être entièrement exclus de la sélection en vue de l'EET. Cette démarche est souvent nécessaire pour des raisons pratiques: il peut simplement ne pas être rentable de tenter une EET dans des secteurs qui ne comportent pas ou très peu d'enfants travailleurs signalés. Les considérations de coûts doivent être mises en balance avec la distorsion qu'introduirait cette exclusion.

Le sous-échantillonnage des secteurs d'ETE

Le sous-échantillonnage des secteurs d'ETE peut être introduit afin de:

- réduire la taille d'échantillon de l'EET; et plus spécifiquement
- concentrer l'échantillon d'EET, autrement dit le confiner à un nombre réduit de secteurs ayant des niveaux de travail des enfants plus élevés.

Les techniques spéciales de sous-échantillonnage visant à atteindre le dernier de ces objectifs sont décrites au chapitre 5.

Élargir les secteurs d'ETE initiaux

Un bref commentaire sera utile à propos de l'option 13. Cette option peut être utilisée quand l'échantillon d'EET nécessite l'inclusion de secteurs additionnels au-delà des secteurs d'ETE. Ce peut être le cas parce que les secteurs de l'échantillon d'ETE ne produisent pas un nombre suffisant de cas d'étude (enfants travailleurs) pour l'EET. Cela peut se produire, par exemple, si la conception initiale de l'ETE était basée sur une attente exagérée du nombre des enfants travailleurs qui seraient découverts. Une autre motivation peut être le renforcement de l'échantillon d'EET par le choix d'un plus grand nombre de cas issus (ou proches) de secteurs d'ETE dont on a constaté qu'ils concentrent des enfants travailleurs. Ces besoins se manifesteront probablement plus fréquemment dans les enquêtes spéciales, portant par exemple sur les lieux où les enfants se rassemblent pour travailler (par ex. porteurs, éboueurs, etc.), mais ils peuvent aussi apparaître dans les enquêtes menées auprès des ménages sur le travail des enfants.

Une manière évidente d'élargir l'échantillon d'EET consiste à y inclure des secteurs additionnels situés dans le «voisinage» des zones de sondage de l'ETE. Diverses techniques peuvent être utilisées pour élargir ainsi l'échantillon tout en maintenant sa nature probabiliste. Quelques procédures simples sont mentionnées dans la section 5.4. Des procédures plus complexes, comme l'«échantillonnage adaptatif par grappes», ont aussi été élaborées.

Dans toutes les autres options énumérées ci-dessus (1 à 12), l'échantillon d'EET est confiné dans les secteurs d'ETE (mais non nécessairement confiné aux ménages constituant l'échantillon d'ETE). Nous pouvons donc considérer que l'option 13 représente le lien le plus lâche entre les deux enquêtes.

2.5.2. La qualité de la sélection assurée par l'ETE

On a noté ci-dessus que deux des facteurs de base déterminant la liaison appropriée entre l'ETE et l'EET étaient (1) l'intervalle séparant les deux enquêtes, et (2) la qualité de la sélection assurée par l'ETE dans l'identification de la présence d'enfants travailleurs dans les secteurs, les ménages ou au niveau individuel.

La question du décalage temporel a déjà été mentionnée à propos des différentes options. Fondamentalement, un intervalle plus long exige des liaisons assurées par des unités plus stables (comme les secteurs, et non les enfants considérés individuellement), et des informations moins détaillées qui devront être reprises d'une enquête à l'autre.

Selon que le questionnaire est plus ou moins détaillé, l'ETE peut varier largement en efficience et en spécificité de sa fonction de sélection pour l'identification du travail des enfants. Quand elle constitue une enquête indépendante ou un module assez détaillé de l'EMO, par exemple, elle a le potentiel d'identifier la présence de travailleurs enfants avec plus de précision, en utilisant une définition appropriée de ce qui constitue le «travail des enfants» selon l'enquête. Dans ce cas, le lien avec la composante EET peut être étroit, ou l'une et l'autre peuvent même être intégrées en une seule opération, à moins que des considérations liées au sous-échantillonnage ou à la complexité de l'ensemble de l'entretien n'amènent à en décider autrement.

Quand l'ETE doit être confinée à un ensemble limité de questions simplifiées, il est probable qu'elle sera moins complète et précise dans l'identification des cas de travail des enfants. Généralement, une version simplifiée du concept de travail des enfants doit être utilisée dans tous les cas. L'information qu'elle fournit sur les cas individuels (enfants) est naturellement moins fiable. Le lien entre les deux composantes n'a donc pas besoin d'être aussi étroit au niveau micro par rapport au cas précédent. Néanmoins, l'information provenant de l'ETE peut être extrêmement utile, par exemple pour déterminer les paramètres d'une conception efficiente de l'EET, tels que la sélection des secteurs en prenant en compte les modèles de concentration du travail des enfants, et la stratification des ménages selon la présence probable ou non d'enfants travailleurs. Ainsi, par exemple, une combinaison des options 11, 8 et 3 de la section 2.5.1 peut en fait être parfaitement appropriée dans de nombreuses situations.

2.6. Les enquêtes sur les enfants et les enquêtes sur les activités des enfants

2.6.1. La portée des enquêtes

Toutes les EET sont centrées sur les détails du travail des enfants ou, plus généralement, sur les activités des enfants en relation avec le travail. Toutefois, certaines enquêtes collectent un plus large éventail d'informations sur les enfants. Dans le tableau 2.2, col. (4) «Portée de l'EET», nous avons distingué les trois types de situation suivants:

1. La majorité des EET sont essentiellement focalisées sur l'étude des conditions et des conséquences du travail des enfants. Dans le tableau, le «champ de l'EET» pour ce type de situation d'enquête a été indiqué comme «TE» (travail des enfants) .

2. Dans un certain nombre de pays, le champ est plus vaste et couvre tous les types d'activités des enfants, incluant les activités économiques et non économiques, l'éducation, les loisirs et même la non-activité. Un grand nombre de ces enquêtes nationales sont en fait appelées «enquêtes sur les activités des enfants» (EAE). Dans le tableau, le champ de l'EET pour ce type de situation d'enquête a donc été indiqué comme «enquête sur les activités des enfants».

3. Dans quelques pays, le champ est encore plus vaste et inclut davantage d'informations générales sur les enfants par-delà leur activité économique et non économique: informations sur la santé des enfants, leurs conditions de logement, etc. On parle alors d'«enquête sur les enfants».

2.6.2. Les incidences sur l'échantillonnage

Ces variations ont des incidences importantes sur l'échantillonnage. Pour les enquêtes centrées sur les enfants travailleurs, option 1, les échantillons d'EET devraient refléter étroitement les modèles de concentration du travail des enfants (chapitre 5). Avec leur champ plus vaste et plus diffus, les enquêtes sur les enfants, option 3, exigeraient une conception semblable à celle de l'ETE (chapitre 4) ou pourraient même incorporer cette dernière. Toutefois, pour ce type d'enquêtes, et particulièrement pour les enquêtes sur les activités des enfants, option 2, l'étude de l'incidence et des conditions du travail des enfants en tant que tel restera probablement un objectif spécial. Une conception de compromis est donc souhaitable, qui couvre à la fois les enfants travailleurs et non travailleurs, mais en donnant davantage de poids aux premiers. Une conception conforme à la méthode du chapitre 6 – avec son compromis dans l'attribution de la probabilité de sélection des secteurs – peut en fait constituer un choix approprié dans de nombreuses situations. Cette conception de compromis exigerait évidemment une opération antérieure de type ETE, afin d'identifier – même avec une précision limitée – le niveau du travail des enfants dans les secteurs de l'enquête.

Nous allons maintenant examiner quelques aspects plus spécifiques de l'échantillonnage pour les enquêtes sur les enfants et sur les activités des enfants.

En premier lieu, nous devons noter qu'abstraction faite de toute information additionnelle collectée sur les enfants, l'*essentiel* de ces différents types d'enquêtes demeure l'étude des conditions et des conséquences du travail des enfants. Dans tous les cas, cette exigence devra donc influencer la taille et la conception de l'échantillon de l'enquête. À cet égard, la population de base est évidemment la *population des enfants travailleurs*.

Dans la conception d'ensemble de l'échantillon, nous avons en fait à considérer trois composantes:

a) La «composante ETE», dont l'objectif premier est d'estimer la proportion (ou le nombre) des enfants engagés dans le travail des enfants.

b) Une «composante supplémentaire», dont l'objectif est d'obtenir un plus large éventail d'informations sur tous les enfants – sur tous les types d'activité des enfants, et éventuellement aussi sur d'autres aspects de la vie de l'enfant.

c) La «composante ETE», dont l'objectif premier est d'obtenir des informations détaillées sur les conditions et les conséquences du travail des enfants.

Dans l'ensemble, on peut s'attendre à ce que les *enquêtes sur les enfants*, plus générales, mettent un peu plus l'accent sur l'objectif (b) que sur les autres objectifs, alors que les *enquêtes sur les activités des enfants* mettront probablement un peu plus l'accent sur (a) et (c). Cela mis à part, les enquêtes particulières de l'un et l'autre types peuvent différer par l'importance relative qu'elles accordent à l'objectif (a) par rapport à l'objectif (c).

Mettre davantage l'accent sur (a) – la mesure de l'ampleur du travail des enfants – rendrait les conditions de la conception plus semblables à celles d'une ETE (chapitre 4). En tout cas, les populations de base pour (a) et (b) sont les mêmes, à savoir la population de tous les enfants. Toutefois, réunir ces deux objectifs exige encore des compromis. Ceux-ci concernent essentiellement des aspects connexes de la complexité des entretiens et de la taille de l'échantillon. Les enquêtes sur l'ensemble des activités des enfants, et notamment les enquêtes plus générales sur les enfants, impliqueront habituellement un entretien beaucoup plus lourd qu'une simple ETE, ce qui aura pour effet de *limiter la taille d'échantillon réalisable*.

En même temps, dans la mesure où l'objectif (b) influe sur le choix, la taille d'échantillon *requise* tend à être aussi plus limitée. Des estimations de l'ampleur du travail des enfants sont fréquemment requises pour des sous-populations et domaines divers dans le pays, et elles exigent habituellement des tailles d'échantillon importantes. Par contre, l'objectif (b) consiste normalement à fournir des informations plus descriptives et détaillées sur les activités des enfants et d'autres aspects de leur vie. Souvent, ce type d'information n'exige pas, en lui-même, une taille d'échantillon aussi importante que dans le cas de (a).

Par ailleurs, l'objectif (c) ne saurait être ignoré, même si l'on y insiste moins. Il milite en faveur d'échantillons plus petits et mieux ciblés. Cette considération est probablement encore plus importante dans une enquête qui met davantage l'accent sur l'objectif (c) par rapport à (a), c'est-à-dire à l'évaluation des conditions et des conséquences du travail des enfants, et non à l'estimation de sa seule ampleur. Ce sera plus probablement la situation que l'on rencontrera dans l'examen de ce qui a été appelé l'activité des enfants, à distinguer des enquêtes sur les enfants. Dans ce cas, l'échantillon devrait être plus petit et mieux ciblé sur des secteurs à forte concentration du travail des enfants.

Une des possibilités consiste à déterminer d'abord (mais en aucune façon exclusivement) la sélection des secteurs sur la base des conditions spécifiques de l'objectif (c), mais à choisir au stade suivant des échantillons représentants tous les enfants (éventuellement avec un sous-échantillonnage des enfants travailleurs) au sein de chaque secteur.

2.6.3. Exemples d'enquêtes sur les enfants et sur l'activité des enfants

L'objectif des exemples suivants, extraits mot pour mot de rapports d'enquête nationaux, est de fournir un tableau plus concret des objectifs et du contenu de ces enquêtes qui, dans leur diversité, ont été menées dans des pays en développement et en transition, généralement dans le cadre des «statistiques sur le travail des enfants».

Encadré 2.7: Exemples d'enquêtes sur les enfants: Ukraine 1999

Les unités d'enquête étaient des enfants de 5 à 17 ans avec un seul parent ou tuteur. …

La période de référence a couvert les trois mois précédant le mois des entretiens. Ceux-ci ont été menés depuis les deux dernières semaines du dernier mois du trimestre jusqu'à la première semaine du trimestre suivant.

Les instruments d'enquête ont été les suivants:

Questionnaire d'enquête sur le travail des enfants à l'intention des parents. Il a été rempli selon les réponses fournies par les parents/tuteurs d'enfants de 5 à 17 ans et comporte 33 questions.

Questionnaire d'enquête sur le travail des enfants à l'intention des enfants. Il a été rempli selon les réponses fournies par des enfants de 5 à 17 ans et comporte 43 questions sur l'éducation et le temps de loisirs, l'activité économique, les conditions de travail, les soins de santé et les travaux ménagers.

Les questionnaires ci-dessus, destinés aux enfants et aux parents/tuteurs, étaient complémentaires, avec certaines questions (par ex. sur les activités des enfants, leurs raisons de travailler, la disponibilité des emplois, l'état de santé et les engagements des ménages) figurant sur les deux questionnaires afin de permettre la comparaison des perceptions des enfants concernant leur travail avec celles de leurs parents. Pour assurer la non-intervention des parents dans les réponses des enfants, ces derniers ont été interrogés en l'absence des parents. Pendant l'analyse des données, les réponses apportées par les parents et les enfants aux mêmes questions ont été comparées pour assurer leur impartialité et la fiabilité des données.

Source: Child Labour In Ukraine 1999, Statistical Bulletin. Bureau International du Travail et State Statistics Committee of Ukraine, 2001.

Encadré 2.8: Exemples d'enquêtes sur les enfants: Cambodge 1996

Objectifs de l'enquête: … recueillir des informations sur les caractéristiques démographiques, les caractéristiques sociales et économiques, l'état de santé des mères et des enfants, les caractéristiques de la main-d'œuvre, le travail des enfants, etc.

Éléments d'information recueillis lors de l'enquête:

Caractéristiques démographiques et économiques.

Pour tous les membres du ménage: relations avec le chef du ménage; âge; sexe; handicap; migration.

Pour les membres du ménage âgés de 5 ans et plus: migration en relation avec l'emploi; éducation; alphabétisme; activité habituelle au cours des 12 derniers mois; activité courante la semaine précédente; profession principale; profession secondaire.

Pour les membres du ménage âgés de 10 ans et plus: situation matrimoniale.

Pour les femmes âgées de 15 à 45 ans: grossesse.

Enfants âgés de 5 à 17 ans (travail des enfants).

Fréquentation scolaire au cours de la semaine passée. Raisons de l'abandon ou de la non-fréquentation scolaire. Raison principale de travailler ou d'avoir un emploi. Âge auquel l'enfant a commencé à travailler. Lieu du travail. Proportion des gains de l'enfant qui sont remis au ménage. Maladie, blessures et autres problèmes de santé des enfants travailleurs. Recrutement d'enfants en vue de les faire travailler ailleurs.

Santé des enfants de moins de 5 ans.

Maladies; traitements administrés; informations sur les enfants de moins de 2 ans; informations sur les enfants de 1 à 2 ans; informations sur les enfants de 6 mois à moins de 5 ans.

Logement et autres particularités du ménage.

Type de bâtiment et année de construction; mode d'occupation et utilisation de l'unité de logement; dépenses du ménage; utilisation de l'eau, de l'éclairage, de combustible pour la cuisine et d'installations sanitaires par le ménage; moyens de transport et éléments de confort possédés par le ménage; accès aux services de base; exploitations agricoles; activités économiques; solvabilité; accidents dans le ménage; … pratiques sanitaires des membres du ménage.

Source: Report on Child labour in Cambodia 1996 (draft). National Institute of Statistics, Phnom Penh, Cambodia, July 1997.).

Encadré 2.9: Exemples d'enquêtes sur l'activité des enfants: Panama 2000

L'objectif général de l'enquête sur le travail des enfants était de mesurer le travail des enfants. Plus spécifiquement, on s'est attaché à déterminer les caractéristiques socio-économiques des logements abritant la population de 5 à 17 ans, à vérifier les conditions de logement de la population des enfants entrant dans la population active, à évaluer les conditions dans lesquelles apparaît le travail des enfants et à obtenir des informations sur les risques et les accidents professionnels qui concernent la population des garçons, filles et adolescents travailleurs. … L'enquête fournit des informations cruciales pour la préparation de politiques spécifiques pour la population de 5 à 17 ans, ainsi que pour le suivi et l'évaluation des programmes qui sont exécutés par différentes agences sociales en vue d'éliminer les pires formes de travail des enfants. Il mérite d'être noté que les données présentées ici permettent une étude détaillée de la population âgée de 5 à 17 ans, au niveau de la population non indigène, en prenant en considération les limites provinciales et les secteurs ruraux et urbains du pays. Dans la province de Panama, elles sont détaillées entre les districts de Panama et de San Miguelito.

Aspects de l'échantillonnage. L'enquête sur le travail des enfants a été menée à l'échelle nationale et a inclus les secteurs indigènes et d'accès difficile, les résidents coutumiers faisant l'objet d'entretiens (enquête de jure). Les segments de personnes interrogées ont été choisis dans un échantillon de segments précédemment connus pour contenir une population âgée de 5 à 17 ans. … Le travail de terrain a été effectué grâce à des entretiens personnels menés à travers le pays dans les logements où une population âgée de 5 à 17 ans avait été décelée, que cette population fût ou non engagée dans le travail.

Source: National report on the results of the child labour survey in Panama 2000. Organisation Internationale du Travail, 2003.

Encadré 2.10: Exemples d'enquêtes sur l'activité des enfants: Belize 2001

Enquête sur l'activité des enfants. Cette EAE a été envisagée pour comptabiliser statistiquement les enfants économiquement actifs, ainsi que pour obtenir des données plus détaillées. Elle était également destinée à fournir l'information nécessaire sur les enfants engagés dans des activités économiques et non économiques, et des caractéristiques démographiques et socio-économiques complètes de tous les enfants d'âge scolaire et des enfants travailleurs: conditions de travail, aspects relatifs à la sécurité et à la santé (l'accent étant mis sur le type, la fréquence et la gravité des blessures ou des maladies) et raisons de travailler. L'enquête visait aussi à identifier les caractéristiques démographiques et socio-économiques des parents de tous les enfants de 5 à 17 ans.

Les enfants travailleurs. L'enquête établit une distinction entre les enfants «économiquement actifs» et «non économiquement actifs» (selon les définitions de l'OIT), les premiers étant les enfants engagés dans toute forme d'activité économique pendant au moins une heure par semaine, les autres les enfants engagés essentiellement

dans le travail domestique non rétribué. Une autre distinction est faite (conformément, en particulier, à la Convention n° 138 de l'OIT concernant l'âge minimum d'admission à l'emploi) pour définir le «travail des enfants», définition applicable à tous les enfants âgés de 5 à 17 ans qui sont économiquement actifs, à l'exception:

- des enfants de 12 à 14 ans engagés dans des travaux légers; et
- des enfants de 15 à 17 ans engagés dans un travail non dangereux (ce qui inclut les travaux légers).

Les «enfants travailleurs» sont ceux qui sont actifs, que ce soit économiquement ou non.

Conclusions et recommandations. En conséquence, un certain nombre de conclusions ont été tirées de l'EAE et des données et informations connexes, ainsi qu'un certain nombre de recommandations, sont présentées dans les secteurs de l'étude ci-dessous:

- conventions internationales et lois et politiques nationales;
- enfants non économiquement actifs;
- enfants économiquement actifs (travailleurs);
- enfants travailleurs et tâches ménagères, scolarité et santé;
- épargne des enfants et contribution apportée aux ménages;
- travail des enfants; et
- pires formes de travail des enfants.

L'EAE a couvert un large éventail de thèmes, dont les caractéristiques du logement, la situation migratoire des ménages, les caractéristiques des enfants vivant éloignés du ménage, les caractéristiques des enquêtés, les caractéristiques démographiques des enfants, la situation migratoire des enfants, l'activité économique des enfants (activité économique actuelle, lieu de travail, employeurs des enfants, gains et horaires de travail au cours de la semaine passée et activité économique habituelle), enfants engagés dans une activité non économique, enfants oisifs, santé et sécurité des enfants qui ont travaillé à un moment donné dans le passé, perception des parents ou des tuteurs des enfants, et questions connexes en direction des enfants. Des caractéristiques générales liées à la situation démographique et socio-économique de la population enquêtée ont également été recueillies. Les résultats sont présentés par sexe, groupe d'âge, secteurs de résidence urbain ou rural, district, groupe ethnique et niveau d'instruction, ainsi qu'en fonction d'autres caractéristiques démographiques et socio-économiques.

Source: Child Labour in Belize: A Statistical Report 2001. International Labour Organization and Central Statistical Office, Government of Belize, 2003.

Encadré 2.11: Exemples d'enquêtes sur l'activité des enfants: Ghana 2001

L'objectif stratégique d'ensemble consiste à fournir des données quantitatives sur les activités des enfants, en incluant la scolarité, les activités économiques et non économiques, ainsi que les travaux ménagers. L'enquête vise aussi à établir une base de données qui servira de référence pour mesurer les progrès accomplis dans l'élimination du problème.

Concepts et définitions:

Travail des enfants. Le nom de l'enquête pourrait suggérer un intérêt exclusif pour les enfants engagés dans des activités économiques, mais le champ de l'enquête va au-delà: elle recherche des informations sur *l'utilisation générale du temps des enfants et les effets sur leur santé, leur éducation et leur développement normal.* De manière générale, toute activité, économique ou non économique, accomplie par un enfant, qui a le potentiel d'affecter négativement sa santé, son éducation, sa moralité et son développement normal relèverait donc du travail des enfants. Dans la détermination de ce qui constitue le travail des enfants, des facteurs comme le nombre d'heures de travail, le type de travail, l'environnement du travail et d'autres ont été pris en compte.

Activité des enfants. «Activité économique» s'applique à tout travail ou activité accomplis pendant une période de référence spécifiée contre un salaire (en espèces ou en nature), un profit ou un gain familial. Dans l'enquête, deux périodes de référence, les 7 derniers jours et les 12 derniers mois, ont été utilisées. Toutes les autres activités sont considérées comme non économiques (tâches ménagères ou travail domestique effectué au sein du ménage, activités volontaires et caritatives, etc.). Puisque les enfants effectuent bien des tâches domestiques dans les ménages de leurs parents/tuteurs, l'enquête a également porté sur les activités des enfants de cette nature.

Le questionnaire destiné aux ménages a recueilli des informations sur les caractéristiques des logements ou des ménages, les caractéristiques socio-démographiques de tous les membres des ménages, l'activité économique, la santé et autres éléments de la situation des enfants. Le questionnaire destiné aux enfants des rues a recueilli des informations sur les caractéristiques socio-démographiques, les dispositions de la vie quotidienne, les origines parentales, les activités économiques, la santé, la sécurité et autres questions liées à la vie dans les rues, et sur l'aide que les enfants des rues attendent de la société et des pouvoirs publics.

Source: The 2001 Ghana Child Labour Survey. Ghana Statistical Service, March 2003.

Encadré 2.12: Exemples d'enquêtes sur l'activité des enfants: Mongolie 2002-03

En Mongolie, l'ENTE 2002-03 a été conduite, conjointement avec l'EMO, pour fournir des estimations fiables du travail des enfants aux niveaux national, urbain et rural, ainsi que par région. L'ENTE a couvert la population des enfants de 5 à 17 ans vivant dans des ménages, cependant que les enfants vivant dans les rues ou les institutions telles que les prisons, les orphelinats ou les centres de protection sociale étaient exclus. Il s'agit là d'une enquête indépendante, et sa taille d'échantillon et sa couverture ont été telles qu'elle a pu fournir des estimations clés assez fiables par région. L'enquête a été conçue pour obtenir des estimations sur le plus grand nombre de variables ou de paramètres possibles, particulièrement en ce qui concerne les activités économiques et non économiques des enfants du groupe d'âge de 5 à 17 ans dans les circonstances habituelles. S'agissant d'une enquête basée sur les ménages au niveau national, elle n'a pas pu appréhender les enfants engagés dans les secteurs des PFTE (pires formes de travail des enfants), ni les enfants laissés à eux-mêmes et vivant dans les lieux publics. …

L'ENTE (autrement dit: EMO + MAE, le module sur l'activité des enfants) est conçue comme une enquête basée sur les ménages qui explore les activités des enfants, définis aux fins de l'ENTE comme étant les personnes âgées de 5 à 17 ans. Les objectifs stratégiques de l'ENTE consistent à générer des données quantitatives sur les activités des enfants (scolarité, activités économiques et non économiques) en Mongolie, et à amorcer le processus d'établissement d'une base de données contenant des informations aussi bien quantitatives que qualitatives sur les activités des enfants. …

L'information requise a été générée dans une approche en deux temps. Tout d'abord, une grande partie des données a été collectée grâce à des entretiens personnels avec les chefs de ménage (ou avec des adultes responsables et informés membres du ménage). On a posé à ces personnes des questions concernant les caractéristiques démographiques et économiques générales de chacun des membres des ménages, y compris les activités des enfants. La seconde partie de l'entretien a été dirigée vers les enfants eux-mêmes et concernait leurs activités (y compris la scolarité), leurs conditions de travail, leurs raisons de travailler, leur perception du travail et leurs plans pour l'avenir.

Le rapport d'enquête fournit des informations complètes sur toutes les activités des enfants du groupe d'âge de 5 à 17 ans qui vivent avec des ménages en Mongolie, et les enfants sont généralement classés comme:

- fréquentant l'école seulement (pas d'autre activité);
- fréquentant l'école et engagés dans une activité économique;
- fréquentant l'école et engagés dans une activité non économique;
- fréquentant l'école et engagés dans une activité économique et dans une activité non économique;
- engagés dans une activité économique seulement;
- engagés dans une activité non économique seulement;
- engagés à la fois dans des activités économiques et non économiques;
- ne fréquentant pas l'école; et
- ne fréquentant pas l'école et non engagés dans des activités économiques et/ou non économiques.

Source: Report on National Child Labour Survey 2002-2003. National Statistical Office of Mongolia, Ulaanbaatar 2004.

Encadré 2.13: Exemples d'enquêtes sur l'activité des enfants: Sri Lanka 1999

Le travail de terrain de l'enquête s'est effectué en quatre phases au cours d'une période allant de novembre 1998 à juin 1999. ... pour la plupart, les enfants n'étaient pas disponibles au moment des entretiens et l'information requise a dû être obtenue à partir d'un enquêté supplétif (très fréquemment d'un parent de l'enfant).

L'équipe de l'enquête a obtenu une très bonne coopération des répondants (les enfants et leurs parents) en ce qui concerne l'information collectée. En effet, le questionnaire ne se limitait pas aux activités économiques des enfants: il était conçu pour obtenir des informations relatives aux activités des enfants sur les plans de l'éducation, des loisirs et des tâches ménagères avant les informations sur les activités économiques, et cet ordre séquentiel des questions a beaucoup aidé les enquêteurs à obtenir la coopération des répondants. Toutefois, la fiabilité des informations fournies par les répondants quant à la durée du temps consacré à chaque activité est discutable, notamment quand l'implication de l'enfant dans une activité particulière est très minime.

Source: Child Activity Survey Sri Lanka 1999. Department of Census and Statistics, Ministry of Finance and Planning, 1999.

Encadré 2.14: Exemples d'enquêtes sur l'activité des enfants: République dominicaine 2000

Une enquête sur les enfants. Cette enquête nationale sur le travail des enfants a été effectuée pour générer des données quantitatives et qualitatives sur les activités économiques, ménagères, scolaires et récréatives des enfants de 5 à 17 ans dans l'ensemble du pays...

La principale population concernée par l'enquête était celle des enfants de 5 à 17 ans résidant sur le sol dominicain. Pour recueillir les données, deux questionnaires différents ont été appliqués: l'un d'eux s'adressait aux ménages et visait à recueillir des données sur le logement, le ménage et les résidents réguliers et visiteurs; l'autre questionnaire était individuel et s'adressait à chacun des enfants de 5 à 17 ans vivant dans les ménages enquêtés.

Les thèmes étudiés au cours de l'enquête étaient les suivants:

Caractéristiques générales de la population de 5 à 17 ans et de ses ménages (population des ménages; répartition par sexe, âge et secteur de résidence; alphabétisme et éducation; caractéristiques des ménages et des logements; composition des ménages avec enfants; logements habités par des enfants; et disponibilité des services de base.

Activités des enfants. Tâches domestiques dans leur propre ménage; activités scolaires; enfants travailleurs; ampleur et caractéristiques des enfants travailleurs; branche d'activité; catégorie dans l'emploi et type de travail; postes et heures de travail; revenus des enfants; scolarité chez les enfants travailleurs.

Facteurs en rapport avec les enfants travailleurs. Raisons de travailler et satisfaction de l'emploi; blessures et maladies liées au travail et sécurité parmi les enfants travailleurs; opinions des parents ou des tuteurs sur le travail des enfants ou des adolescents; opinions des enfants travailleurs; contexte des ménages des enfants.

Source: Report on the results of the National Child Labour Survey in the Dominican Republic, 2000. International Labour Office and State Department of Labour of the Dominican Republic, 2004

Encadré 2.15: Exemples d'enquêtes sur l'activité des enfants: Costa Rica 2003

L'Enquête à objectifs multiples auprès des ménages (EOMM) est une source précieuse d'informations sur différentes questions relatives aux ménages du pays et à leurs membres. Le module de l'enquête relatif à l'emploi, centré sur la main-d'œuvre et ses caractéristiques (emploi, chômage, sous-emploi et revenus), est particulièrement important. L'EOMM fournit aussi des informations complémentaires sur les caractéristiques démographiques, socio-économiques et éducatives des individus et des ménages. L'EOMM a été conduite en juillet de chaque année depuis 1987. L'ajout d'un module sur le travail des enfants en 2002 a fait partie de l'élaboration du projet d'enquête sur le travail des enfants et de développement d'une base de données.

L'objectif du module était de collecter des informations sur l'ampleur et les caractéristiques de la participation des filles et des garçons âgés de 5 à 17 ans à des activités économiques, récréatives, éducatives et domestiques, ainsi que sur les principales caractéristiques démographiques et socio-économiques des mineurs travailleurs.

Source: National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica, 2003. Bureau International du Travail, 2004.

2.7. Les enquêtes sur les unités autres que les ménages

On peut noter les points suivants concernant les enquêtes dans les établissements et les écoles dans leurs relations avec les enquêtes auprès des ménages.

On peut distinguer deux types d'études impliquant des établissements. Le premier consiste en des enquêtes représentatives de la population des établissements qui emploient des enfants. Ici, les établissements ciblés par l'enquête peuvent être choisis à partir des répertoires ou des listes disponibles, dont ceux qui concernent les associations de producteurs et les coopératives, ou à partir de listes établies au cours d'une enquête au niveau communautaire ou d'une enquête auprès des ménages sur le travail des enfants. Les listes peuvent aussi être basées sur des enquêtes locales dans le secteur à étudier, par exemple auprès d'informateurs tels que les syndicats, les agences gouvernementales, les ONG, les dirigeants communautaires, les groupes religieux ou les associations caritatives.

Le second type d'études implique des unités liées, extérieures aux ménages, qui sont introduites dans l'enquête sur la base d'une liaison avec un ménage ou une personne de l'échantillon.

Cela concerne les enquêtes spéciales impliquant divers types d'unités – écoles, établissements ou autres lieux où travaillent des enfants, et éventuellement autres types d'institutions non résidentielles influant sur les enfants travailleurs – qui sont liées aux enquêtes sur le travail des enfants basées sur les ménages. Par exemple, au Bangladesh (2002-2003) et au Cambodge (1996), des questionnaires spéciaux pour les employeurs ont

été utilisés. Dans ce type d'enquêtes liées aux enquêtes basées sur les ménages, les unités extérieures aux ménages incluses dans l'étude sont normalement identifiées grâce aux enfants travailleurs dénombrés dans l'enquête principale basée sur les ménages. En général, aucune nouvelle opération d'échantillonnage n'est impliquée dans la sélection de ces unités, à l'exception peut-être d'un sous-échantillon d'enfants de l'enquête principale pour lesquels les unités liées extérieures aux ménages doivent être dénombrées. En outre, la portée et la couverture de l'enquête liée sur les unités extérieures aux ménages sont habituellement assez restrictives ou sélectives. Pour ces raisons, il n'est généralement pas approprié de considérer les informations provenant de celles-ci comme représentatives de la population des unités extérieures aux ménages concernés (par ex. comme représentatives de la population de toutes les écoles, de tous les lieux de travail employant des enfants, etc.). Une représentation adéquate de ces populations exigerait généralement des échantillons d'une taille et d'une conception très différentes de celles de l'ensemble d'unités obtenu simplement sur la base d'une association à un échantillon d'enfants. Il est plus approprié d'interpréter les données résultant de ces échantillons restreints d'unités extérieures aux ménages comme une source de variables additionnelles qui peuvent être attachées à d'autres informations sur l'enfant concerné. Dans la mesure où l'échantillon d'enfants - en liaison avec lequel des unités extérieures aux ménages sont introduites dans l'enquête - est représentatif de la population d'enfants, les résultats peuvent être généralisés à cette population.

Un exemple est donné par les enquêtes sur les employeurs et les établissements identifiés par association à des enfants travailleurs dans un échantillon d'enquête sur le travail des enfants basée sur les ménages. De même, les enquêtes basées dans les écoles sont généralement du type mentionné ci-dessus et portent sur les enseignants et les écoles identifiés par association avec les enfants d'un échantillon d'enquête sur le travail des enfants basée sur les ménages. Comme on l'a indiqué, dans ce type d'études, l'échantillon n'est pas conçu pour être représentatif de la population totale des enseignants ou des écoles: ces unités sont plutôt interrogées pour obtenir des variables additionnelles sur l'échantillon d'enfants associés à ces unités[8].

2.8. Enquête sur les activités des jeunes en Afrique du Sud 1999: description et commentaires

L'*Enquête sur les activités des jeunes 1999* (SAYP) a été la première enquête de ce type en Afrique du Sud.

Les informations suivantes sont extraites du *Rapport national sur les activités des enfants en relation avec le travail*, Agence sud-africaine de la statistique (2001).

[8] Nous avons souligné ci-dessus que, de manière générale, on ne peut pas s'attendre dans la pratique à obtenir un échantillon représentatif d'institutions telles que les écoles ou les établissements employant des enfants par association avec les enfants d'un échantillon d'enquête basée sur les ménages. Bien entendu, il est en principe possible d'y parvenir, à condition que cet échantillon soit «suffisamment important, à large base et à couverture satisfaisante». Ce qui est en jeu ici est l'échantillonnage indirect des institutions. Cet échantillonnage s'effectue dans une situation où nous n'avons aucun accès direct à la population cible (institutions) pour la sélection de l'échantillon, mais seulement à des populations en relation avec celle-ci (enfants travailleurs dans des ménages privés). On dispose de procédures permettant de produire des estimations pour la population d'unités sélectionnées indirectement de cette manière. Pour des faits nouveaux et récents, voir Lavallée, P. (2007), *Indirect sampling,* Springer.

2.8.1. Objectifs et structure

Objectifs de l'enquête

Les objectifs de l'enquête étaient les suivants:

1. produire des données statistiques complètes sur les activités de travail des jeunes personnes au niveau national;
2. créer une base de données spéciale sur les activités de travail des jeunes personnes en Afrique du Sud, qui sera mise à jour au fur et à mesure de la mise à disposition d'informations statistiques récentes grâce à de nouvelles enquêtes;
3. fournir une analyse exhaustive de la situation des enfants travailleurs du pays en identifiant les paramètres majeurs, les groupes et modèles prioritaires, l'ampleur et les déterminants du travail des enfants, les conditions et effets du travail, etc.;
4. examiner de près des facteurs tels que l'activité économique, les tâches ménagères et activités d'entretien à l'école excessives qui peuvent influer sur la capacité des jeunes de fréquenter l'école ou de s'engager dans d'autres activités d'enfance;
5. diffuser aussi largement que possible les résultats d'une enquête nationale sur les activités des jeunes, en particulier dans les secteurs où ces activités sont le plus intensives;
6. formuler un module sur les activités des jeunes qui sera rattaché à l'une des phases de la nouvelle enquête semestrielle sur la main-d'œuvre récemment introduite, une fois que celle-ci sera pleinement élaborée et deviendra opérationnelle;
7. renforcer la capacité de Stats SA à conduire plus régulièrement dans l'avenir des enquêtes nationales sur ces activités.

Structure de l'enquête

L'enquête a permis de recueillir des informations détaillées en deux phases.

Au cours de la phase 1, des entretiens ont été menés pour déterminer l'ampleur des activités de travail chez les enfants et pour recueillir des informations démographiques générales sur les ménages afin de permettre l'analyse de liens existant entre ces facteurs et les enfants travailleurs. Le but essentiel de cette phase était d'*identifier les ménages dans lesquels un enfant au moins avait été engagé dans des activités de travail des enfants*. Elle comportait aussi des questions concernant les facteurs qui peuvent inciter ou non les enfants à s'engager dans le travail, pour permettre les comparaisons entre les ménages dans lesquels existe le travail des enfants et les autres. Les questions du premier questionnaire s'adressaient à un adulte responsable, de préférence de sexe féminin, qui s'occupait des enfants dans le ménage.

Au cours de la phase 2, des entretiens complémentaires ont été organisés auprès des ménages de la phase 1 qui avaient au moins un enfant engagé dans une «activité en relation avec le travail», selon la définition donnée ci-dessous. Une caractéristique centrale de la conception de l'enquête était donc que le questionnaire de la première phase était administré dans tous les ménages choisis et que les réponses à certaines

questions étaient utilisées pour trier ou sélectionner les ménages en vue de la seconde phase. Sous réserve d'un sous-échantillonnage dans certains cas, tous les ménages possédant certaines caractéristiques étaient choisis pour la seconde phase.

Les questions de sélection ont été dirigées vers le principal répondant au questionnaire de première phase. Un ménage était considéré comme ayant un enfant engagé dans une activité en relation avec le travail si l'un des enfants du ménage:

a) avait été engagé, à un moment quelconque au cours des 12 derniers mois, dans une des activités économiques suivantes *contre un salaire, un profit et/ou un gain familial:*

- diriger n'importe quel type d'entreprise, grande ou petite, pour l'enfant lui-même;
- apporter une aide non rémunérée à une entreprise familiale;
- aider à des activités agricoles sur la parcelle familiale, le potager, le pâturage communal ou le kraal;
- capturer ou cueillir toute espèce de poisson, de crustacé, de coquillage, d'animal sauvage ou toute autre nourriture, pour la vente ou la consommation familiale;
- effectuer tout type de travail contre une rémunération, un salaire ou tout paiement en nature; ou
- mendier de l'argent ou de la nourriture en public;

et/ou

b) avait été engagé régulièrement pendant une heure par jour ou plus dans une ou plusieurs des activités suivantes:

- tâches ménagères au sein de leur ménage;
- aller chercher du bois et/ou de l'eau, ou effectuer tout travail domestique non rémunéré; ou
- aider à nettoyer et à améliorer l'école.

D'après ce qui précède, il apparaît clairement que deux critères distincts ont été utilisés concernant les différents types d'activités économiques:

1. Dans le cas des activités économiques effectuées contre un salaire, un profit et/ou un gain familial, un *très large* critère de sélection a été utilisé. Chaque ménage dont un enfant de 5 à 17 ans avait été engagé dans ces activités, à un moment quelconque dans les 12 mois précédents (et ne serait-ce qu'une seule fois), a été sélectionné pour la seconde phase. Aucun filtre basé sur la durée n'a été appliqué. Par conséquent, un ménage a été choisi même quand un enfant avait participé à ce type d'activité pendant une très courte période (par ex. une demi-heure) au cours des 12 derniers mois.

2. Quand il s'agissait d'aller chercher du bois et/ou de l'eau, ou d'effectuer tout travail domestique non rémunéré, *un filtre plus restrictif* a été appliqué. Un ménage était

choisi à cause de la participation d'un enfant à ces activités seulement s'il y avait été engagé de manière régulière pendant au moins une heure par jour.

Le but essentiel de la phase 2 était d'explorer plus en détail le travail des enfants. Des questions plus approfondies sur la nature du travail effectué par les enfants étaient posées à un adulte du ménage, ainsi qu'aux enfants impliqués dans ces activités.

Commentaires

1. Selon la terminologie de ce manuel, la phase 1 est une enquête indépendante sur le travail des enfants (ETE), alors que la phase 2 est une enquête liée sur les enfants travailleurs (EET). Dans la première, la population de base englobe tous les enfants d'une fourchette d'âge spécifiée, et l'objectif principal est de mesurer la proportion de ceux qui sont engagés dans le travail des enfants. Dans l'EET, la population de base se limite aux enfants travailleurs, et l'objectif est d'étudier ses caractéristiques et ses circonstances – ce qui n'exclut pas d'obtenir des informations semblables sur un échantillon d'enfants non travailleurs à titre de comparaison.

   La liaison entre les deux enquêtes ETE et EET peut revêtir des formes différentes, allant de l'échantillonnage complètement indépendant (EET «autonome») à une extrémité, à l'EET menée sur un ensemble d'enfants considérés individuellement, avec des caractéristiques spécifiées déterminées à partir du dénombrement de l'ETE, à l'autre extrémité. Dans le cas présent, la liaison est très étroite: l'EET est appliquée à un ensemble de ménages comportant un ou plusieurs enfants avec des caractéristiques spécifiées, tels qu'ils ont été identifiés à partir du dénombrement de l'ETE.

2. L'information sur les ménages sans enfants âgés de 5 à 17 ans et, plus important, sur les ménages qui comportent seulement des enfants non travailleurs dans ce groupe d'âge est obtenue seulement au cours de la phase 1. L'information détaillée de la phase 2 est limitée (1) au groupe des enfants engagés dans une «activité en liaison avec le travail des enfants», plus (2) les enfants qui, tout en n'étant pas eux-mêmes engagés dans une activité en relation avec le travail, vivent dans un ménage où un autre enfant est ainsi engagé. Les comparaisons entre ces deux groupes fournissent une base pour l'analyse des conséquences du travail des enfants. On peut toutefois faire observer que l'information est incomplète – et peut en fait conduire à des distorsions de résultats dans l'analyse des déterminants du travail des enfants. En effet, les deux groupes étudiés sont limités aux «ménages à potentiel de travail des enfants» – identifiés dans l'enquête comme des ménages actuellement sujets à la présence de travail des enfants. Les enfants vivant dans des ménages «sans potentiel de travail des enfants» en ce sens (autrement dit, les ménages qui ont des enfants, mais dont aucun n'est engagé dans une activité en relation avec le travail) ne sont pas examinés, ce qui risque de limiter fortement l'analyse, au niveau des ménages, des déterminants du travail comme les revenus, la consommation et la situation de l'emploi du ménage, ainsi que d'autres caractéristiques du ménage.

L'information requise pour ce type d'analyse peut provenir de l'un ou l'autre des deux sources suivantes:

1. Collecte de l'information requise, à un niveau de détails suffisant, sur tous les types de ménages comportant des enfants et sur toutes les catégories d'enfants dans ces ménages au cours de la phase 1, et/ou

2. inclusion dans la phase 2 d'un échantillon de ménages comportant des enfants, mais dont aucun n'est engagé dans une activité en relation avec le travail.

La conception présente ne fournit pas suffisamment d'informations de ce type.

D'un autre côté, l'enquête est avant tout centrée sur les enfants travailleurs, et une grande partie des ressources devrait donc leur être consacrée. La plupart des objectifs d'enquête, dont la liste figurant dans le rapport d'enquête, concernent la situation et les conditions de vie des enfants travailleurs. Bien entendu, les enfants non travailleurs doivent aussi être étudiés pour fournir un «groupe de contrôle», mais il n'est pas économique – dans une enquête sur le travail des enfants, à distinguer d'une enquête générale sur les enfants – de consacrer une part trop importante des ressources à cet aspect.

Population cible et couverture de l'enquête

La SAYP de l'Afrique du Sud était une enquête basée sur les ménages, et les données ont été collectées à l'occasion d'entretiens individuels avec les enquêtés. La population cible de la phase 1 était composée de tous les enfants de moins de 18 ans qui résidaient normalement dans un ménage privé à l'intérieur du pays. L'enquête excluait donc les enfants qui ne vivaient pas dans un ménage, par exemple les enfants de la rue et les enfants vivant en permanence dans des institutions. Selon les règles concernant le champ de l'enquête, tous les enfants qui étaient des *résidents habituels* devaient y être inclus, même s'ils n'étaient pas présents au moment de l'enquête. Cela signifie que la plupart des pensionnaires étaient inclus dans le ménage de leurs parents. Les activités des enfants de moins de 5 ans n'ont pas été examinées dans cette enquête, car on a estimé qu'ils étaient trop jeunes pour répondre aux questions appropriées.

Un *ménage* a été défini comme consistant en «une seule personne ou un groupe de personnes ayant ou non des liens de parenté qui vivent habituellement ensemble pendant au moins quatre nuits par semaine, qui prennent leur repas ensemble et qui partagent leur ressources». Si un membre habituel du ménage a été absent pendant plus de 30 jours, il n'est plus considéré comme faisant partie du ménage. Les invités et les visiteurs qui ont séjourné 30 jours ou plus sont comptés parmi les membres du ménage. Un ménage peut occuper plus d'une structure. Les personnes qui occupent la même unité d'habitation, mais qui ne partagent pas leurs repas ou d'autres éléments essentiels de la vie, sont considérées comme des ménages séparés. Un travailleur domestique vivant dans un logement séparé réservé aux travailleurs domestiques, ou qui est payé en espèces par le ménage principal (même s'il prend la plupart de ses repas avec ce ménage), est considéré comme un ménage séparé.

L'enquête a été menée dans tout le pays, à la fois dans les secteurs urbains et ruraux des neuf provinces. La population échantillonnée excluait toutes les personnes emprisonnées, les patients hospitalisés, les internes résidant dans des pensionnats et les personnes habitant des pensions de famille, les hôtels et les foyers de travailleurs. Toutefois, les familles vivant dans ces foyers étaient prises en compte. Les ménages composés d'une seule personne ont été éliminés par sélection dans tous les secteurs avant le tirage au sort de l'échantillon: ces ménages ne présentaient pas d'intérêt, puisqu'ils ne comportaient aucun membre de la population cible (enfants âgés de 5 à 17 ans).

Commentaires

1. On notera que la population cible est constituée par l'ensemble des enfants, et non par la totalité de la population résidant dans les ménages privés. Cela signifie que les ménages ne contenant pas d'enfants ne présentent, en tant que tels, aucun intérêt dans cette enquête. Ils peuvent être exclus de l'enquête si l'information a été collectée au cours de l'opération de listage des ménages pour les identifier avant la sélection de l'échantillon. Sinon, ils doivent être soumis au processus d'échantillonnage, et identifiés et éliminés au cours des entretiens sur le terrain.

   La présente enquête exclut une majorité des ménages sans enfant en identifiant et en excluant les ménages comportant une seule personne au stade du listage. Une telle stratégie est rentable.

2. Dans les enquêtes sur le travail des enfants, il est généralement préférable d'utiliser – comme on l'a fait dans cette enquête – la définition de jure (plutôt que de facto) du champ couvert par l'enquête. Cela facilite la couverture des membres des ménages qui peuvent être temporairement éloignés pour des raisons de travail.

3. Si l'obtention d'estimations sur l'activité économique de la population tout entière avait également fait partie des objectifs, il aurait été nécessaire d'inclure dans l'échantillon les ménages sans enfants.

2.8.2. Contenu

Questionnaires

Deux questionnaires ont été élaborés pour l'enquête, un pour chaque phase.

Phase 1

Le questionnaire de la première phase a été utilisé à des fins de sélection et couvrait l'information de base sur les caractéristiques des ménages et tous les membres des ménages. Il a été demandé à un adulte responsable des enfants du ménage (généralement une personne de sexe féminin) de fournir l'information requise dans cette phase: cela incluait l'information démographique de base sur les ménages, comme l'âge des membres du ménage, les relations familiales entre les membres du ménage, le niveau d'instruction le plus élevé, et les activités économiques et non économiques des enfants. Plus spécifiquement, le questionnaire portait sur les points suivants:

- conditions de vie du ménage, incluant le type de logement, le combustible utilisé pour la cuisine, l'éclairage et le chauffage, la source d'eau à usage domestique, la propriété foncière, l'occupation des terres et les cultures;
- information démographique sur les membres du ménage, c'est-à-dire sur les adultes comme sur les enfants - questions portant sur l'âge, le sexe, la situation matrimoniale, le niveau d'instruction et les relations de parenté de chaque membre du ménage;
- la migration du ménage dans les deux années précédant l'enquête;
- le revenu du ménage;
- la fréquentation scolaire des enfants de 5 à 17 ans;
- pour les enfants de 5 à 17 ans, des informations sur les activités économiques et non économiques dans les 12 mois précédant l'enquête.

Phase 2

L'analyse des activités économiques de travail des enfants dans le rapport d'enquête principale a été établie à partir des questionnaires administrés dans les ménages choisis après le processus de sélection de la première phase. Le questionnaire de la seconde phase a été administré au sous-ensemble échantillonné de ménages dans lesquels un enfant au moins avait participé à une forme quelconque de travail au cours de l'année précédant l'entretien (avec un sous-échantillonnage des ménages s'ils étaient trop nombreux dans l'unité primaire d'échantillonnage - UPE - de la phase 1). Une fois encore, il a été demandé à un adulte responsable des enfants du ménage (généralement une personne de sexe féminin) de fournir des informations sur la situation économique et le revenu des adultes des ménages, ainsi que certaines informations de base sur les enfants (avec davantage de détails sur les plus jeunes enfants âgés de cinq à neuf ans). Il a ensuite été demandé aux enfants de fournir eux-mêmes l'essentiel des informations sur leurs activités, en limitant les recoupements avec les réponses des adultes. L'analyse du rapport d'enquête principale est basée, quand elle est disponible, sur les informations fournies par les enfants eux-mêmes. On considère que dans une grande majorité des cas, ces derniers peuvent répondre eux-mêmes à ces questions.

Un ou plusieurs adultes de chaque ménage ont répondu aux questions suivantes:

- situation de l'emploi de tous les adultes du ménage âgés de 18 ans ou plus;
- précisions sur le type de travail dans lequel les adultes étaient engagés;
- revenus de chaque adulte au cours des 12 derniers mois;
- type d'activité en relation avec le travail dans laquelle était (éventuellement) engagé chaque enfant du ménage âgé de 5 à 17 ans;
- raisons qu'avaient le/les enfant(s) de s'engager dans ces activités;
- fréquentation scolaire et problèmes rencontrés à l'école concernant les enfants âgés de 5 à 9 ans;
- sécurité et santé, maladies et blessures liées aux activités en relation avec le travail, concernant les enfants âgés de 5 à 9 ans.

Chaque enfant du ménage âgé de 5 à 17 ans devait répondre aux questions suivantes:

- engagement de l'enfant dans des activités en relation avec le travail au cours de l'année passée et au cours des 7 derniers jours, et type d'activité (le cas échéant);
- précisions sur le type de travail, le secteur et la profession, chez les enfants engagés dans des activités économiques contre un salaire, un profit ou un gain familial économique;
- périodes de l'année et heures de la journée pendant lesquelles ont travaillé ceux qui participaient à ces activités;
- raisons de s'engager dans ces activités;
- conditions dans lesquelles a eu lieu le travail, y compris l'éventualité de harcèlement ou de violence sexuelles (si l'enfant était salarié);
- revenu acquis, et proportion des gains remise aux adultes du ménage; sécurité et santé, maladies et blessures, en relation avec l'activité économique (question posée seulement aux enfants de 10 à 17 ans);
- situation de recherche de travail de l'enfant;
- activité principale de l'enfant;
- fréquentation scolaire et (si l'enfant fréquente l'école) difficultés rencontrées à l'école;
- raisons de manquer l'école ou de ne pas la fréquenter.

Commentaires

1. Les informations recueillies au cours de la phase 1 sur les caractéristiques des ménages et les enfants considérés individuellement dans les ménages où personne n'est engagé dans le travail des enfants sont très limitées (ainsi, la situation de la main-d'oeuvre ne semble pas avoir été incluse). Il aurait été utile de disposer d'autres informations du même type pour analyser les déterminants du travail des enfants.

2. Dans le modèle actuel, les questions relatives à l'activité économique des enfants sont cruciales, car ces questions identifient les cas à dénombrer en phase 2. Diverses raisons contribuent à la tendance au sous-dénombrement des enfants travailleurs dans les enquêtes générales sur les ménages.

3. Ainsi, l'échantillon de l'EET (c'est-à-dire de la phase 2) est conditionné par la réussite de la sélection visant à identifier les ménages qui comportent des enfants travailleurs, et cela essentiellement en termes d'activité économique habituelle (y compris les travaux ménagers substantiels), plutôt que de définition plus complète du travail des enfants. Bien entendu, cette pratique est très répandue.

   Exclure entièrement les ménages qui ne signalent pas de travail des enfants empêche d'évaluer la qualité de l'opération de sélection. Les points positifs sont que (i) accomplir des tâches ménagères substantielles a aussi été inclus comme critère, et que (ii) dans les ménages identifiés, tous les enfants – et non pas

seulement ceux qui ont précédemment été identifiés comme économiquement actifs – ont été étudiés dans la phase 2 (la composante EET).

Calendrier et périodes de référence

L'enquête a été menée en juin et au début de juillet 1999. Les périodes de référence suivantes ont été utilisées:

- Pour la plupart des questions concernant les activités économiques des enfants effectuées en échange d'un salaire, d'un profit ou d'un gain économique familial (définies ci-dessous), la période de référence s'étendait sur les 12 mois précédant l'entretien.
- Quelques questions ont été posées sur les activités économiques actuelles des enfants effectuées en échange d'un salaire, d'un profit ou d'un gain économique familial. La période de référence pour cette section couvrait les sept jours précédant l'entretien.
- Pour la plupart des questions concernant les activités non économiques, le ramassage du bois et/ou le transport de l'eau et le travail domestique non rémunéré (défini ci-dessous), la période de référence couvrait les sept jours précédant l'entretien.
- Certaines questions concernant les activités économiques et non économiques mentionnaient les activités «habituelles», sans indiquer de période de référence.
- Quelques questions concernant les risques pour la santé et la sécurité (comme les conséquences d'une blessure) impliquaient une période de référence ouverte. En voici deux exemples: «Avez-vous déjà été blessé en effectuant une de ces activités économiques en échange d'un salaire, d'un profit ou d'un gain économique familial ?» et «Avez-vous déjà eu à effectuer un travail physique important ?»

Commentaire

Le présent travail d'enquête sur le terrain a été conduit pendant une période relativement brève de quelques semaines. Un tel arrangement est fréquent et peut présenter des avantages sur le plan du coût de l'enquête et de la qualité des données collectées. Toutefois, le travail de terrain devra être prolongé si l'on doit contrôler ou capter des effets saisonniers importants.

Tester la méthodologie et les questionnaire: test de sélection

Il était nécessaire d'obtenir une idée globale de la proportion de ménages qui comportent des enfants engagés dans des activités en relation avec le travail, autrement dit un «taux de présence». Cette information était nécessaire pour la conception et la planification du test pilote ainsi que pour l'enquête principale. Elle a été obtenue grâce à un test de sélection organisé en janvier 1999. Ce test a été conduit dans 28 secteurs de dénombrement (SD), soit: six secteurs urbains formels; six secteurs urbains informels; huit secteurs d'exploitations agricoles commerciales; et huit autres secteurs ruraux (essentiellement des anciens secteurs appartenant aux «homelands»). Les SD du test

de sélection ont été délibérément choisis pour représenter des conditions diverses. On a obtenu les taux de présence suivants:

- 11 pour cent concernant les activités économiques effectuées en échange d'un salaire, d'un profit ou d'un gain économique familial, c'est-à-dire que dans 11 pour cent des ménages, un enfant au moins avait été engagé dans ce type d'activités au cours des 12 mois précédant l'entretien; et
- 23 pour cent concernant le travail domestique non rémunéré, le fait d'aller chercher du bois et/ou de l'eau, les tâches ménagères et le travail scolaire (en excluant les ménages concernés par le premier taux de présence). Cela signifie que dans 23 pour cent des ménages, un enfant au moins avait été régulièrement impliqué dans ces activités pendant au moins une heure par jour, et qu'aucun enfant n'avait participé à des activités économiques effectuées en échange d'un salaire, d'un profit ou d'un gain économique familial. Ces résultats indiquaient que la stratégie d'échantillonnage prévue (basée sur un taux de présence minimum de 20 pour cent pour les deux ensembles d'activités réunis) avait largement atteint sa cible. Il a été convenu d'augmenter la taille de l'échantillon dans les secteurs urbains informels et dans les secteurs d'agriculture commerciale. Le test de sélection a également indiqué que, puisque le taux de présence était souvent supérieur à 20 pour cent un sous-échantillonnage serait nécessaire au cours de la seconde phase de l'enquête pour limiter la taille de l'échantillon obtenu.

Commentaire

Dans de nombreuses situations, on dispose d'informations insuffisantes sur la proportion d'enfants qui sont engagés dans le travail des enfants. Ces informations jouent un rôle crucial dans la planification et le contrôle de la taille d'échantillon des enquêtes sur les enfants travailleurs. L'enquête menée en Afrique du Sud fournit un excellent exemple de bonnes pratiques. Les «tests de sélection» du type décrit ci-dessus peuvent être utiles pour améliorer les informations disponibles pour la conception de l'enquête. Même dans ces conditions, il est parfois nécessaire d'ajuster les taux de sélection de l'échantillon à un stade ultérieur pour tenter de parvenir à la taille d'échantillon prévue pour l'EET.

2.8.3. Aspects de l'échantillonnage

Base de sondage et stratification

La base de sondage utilisée pour la sélection des secteurs était basée sur les secteurs de dénombrement (SD) du recensement de la population de 1996. Pour ce recensement, tout le pays avait été divisé en quelque 86 000 SD, regroupés en seize types de SD.

Les SD ont été explicitement stratifiés par province. À l'intérieur de chaque province, ils ont été à nouveau stratifiés selon les quatre types de SD obtenus en regroupant les 16 types utilisés dans le recensement de la population de 1996. Ces quatre types, définis ci-dessous, étaient: urbain formel, urbain informel, autres ruraux, et agriculture commerciale. Les SD constitués uniquement d'institutions ont été exclus.

Les secteurs urbains sont ceux qui ont officiellement été déclarés comme tels. Ils incluent les villes, les cités et les secteurs métropolitains. Les secteurs urbains sont divisés en:

- *secteurs urbains formels,* constitués essentiellement de logements composés de matériaux de construction formels, comme la brique; et
- *secteurs urbains informels,* constitués essentiellement d'habitations précaires faites de matériaux informels tels que le carton ou la tôle ondulée.

Les secteurs non urbains incluent les exploitations agricoles commerciales, les petits établissements, les villages, les terres traditionnelles et autres zones rurales, qui sont plus éloignées des villes et des cités. Les secteurs non urbains sont divisés en:

- *secteurs d'agriculture commerciale,* comportant des fermes qui vendent l'essentiel de leur production à des fins lucratives;
- *autres secteurs ruraux,* composés de la plupart des secteurs non urbains autres que les secteurs d'agriculture commerciale. On les trouve essentiellement mais non exclusivement dans les anciens «homelands».

La plupart des SD comptaient moins de 100 ménages. Des petites parcelles comportant au moins 100 ménages chacune, appelées unités primaires d'échantillonnage (UPE), ont été délimitées à partir des SD. Une UPE était censée être: soit un seul secteur de dénombrement du recensement de 1996, s'il comptait au moins 100 ménages; soit une combinaison de secteurs de dénombrement adjacents, s'ils comptaient moins de 100 ménages. Cette disposition répondait à la nécessité d'avoir un minimum de 100 ménages dans chaque UPE.

Commentaire

Un tel regroupement de secteurs exhaustif et uniforme dans la totalité du plan ne devrait être introduit que s'il peut s'effectuer de manière économique et sans trop de difficultés dans le cadre d'une opération administrative. Dans le cas contraire, on recherchera des solutions de substitution moins onéreuses. Plusieurs procédures de ce type sont proposées dans les manuels d'échantillonnage.

Système d'échantillonnage

Échantillonnage en deux phases

La collecte de données devant s'effectuer en deux phases, l'échantillon de l'enquête a lui aussi été tiré au sort en deux phases.

Au cours de la première phase, 900 UPE composées au moins de 100 ménages chacune ont été choisies en utilisant des techniques d'échantillonnage probabiliste. Parmi ces UPE, 579 étaient situées en secteur urbain (372 en «urbain formel»; 207 en «urbain informel»), et 321 dans des secteurs non urbains (180 en «agriculture commerciale»; 141 en «autres ruraux»).

La taille de l'échantillon a été déterminée disproportionnellement aux strates explicites en utilisant la méthode de la racine carrée (échantillon proportionné à la racine carrée de la taille de population de la strate). Le nombre d'UPE attribuées a été systématiquement choisi avec probabilité proportionnelle à la taille dans

chaque strate explicite (la taille étant déterminée en fonction du nombre de ménages inclus dans une UPE). À l'intérieur des strates, les SD ont été ordonnés par circonscription judiciaire, et à l'intérieur de celle-ci par type de SD, puis par type de secteur – ce qui assurait donc une «stratification implicite».

Listage

Avant de sélectionner les ménages de la phase 1, toutes les habitations et tous les ménages au sein des UPE échantillonnées ont été listés avant le travail sur le terrain. L'unité finale d'échantillonnage était l'unité d'habitation. Cette unité était formellement définie comme suit:

> «Est une *unité d'habitation* toute structure dans laquelle peuvent vivre des personnes. Un ménage peut en occuper une ou plusieurs. À l'inverse, une seule unité d'habitation peut être occupée par plusieurs ménages. De plus, toute structure ou partie d'une structure qui est inoccupée mais dans laquelle il est possible de vivre est également listée comme unité d'habitation séparée. Une unité d'habitation peut être une maison, un appartement, une case, une péniche, etc., dans lesquels vit ou peut vivre un ménage.»

Les unités d'habitation spéciales, sans occupation privative par un ménage, n'ont pas été considérées comme des unités d'habitation et n'ont donc pas été prises en considération pour la sélection des ménages. Il s'agit des espaces destinés aux patients dans les hôpitaux, aux détenus des prisons et des maisons de redressement, aux personnes vivant dans des pensions et des hôtels, aux pensionnaires de résidences pour personnes nécessitant des soins particuliers (ex.: handicapés, personnes âgées, etc.) et aux élèves internes des pensionnats, à condition que leurs repas proviennent d'une cuisine commune. Les foyers de travailleurs possédant une cuisine commune ont également été traités comme des unités d'habitation spéciales. Toutefois, si une partie de ces foyers hébergeait des familles vivant de manière autonome, les ménages qui y vivaient ont été listés et inclus dans la sélection. En Afrique du Sud, et notamment dans les secteurs ruraux et les établissements informels, les unités d'habitation n'ont pas nécessairement d'adresse. Il était donc important de procéder à un listage complet des ménages de chaque UPE sélectionnée pour s'assurer qu'un même ménage pouvait être repéré sur une carte et sur le terrain, et qu'il pouvait recevoir une nouvelle visite.

Sélection des unités finales

Les unités d'échantillonnage finales étaient les habitations occupés. Lorsqu'on trouvait des ménages multiples dans une unité d'habitation sélectionnée, ils étaient tous interrogés.

Dans la phase 1, au sein de chaque UPE des secteurs urbains, 25 ménages ont été interrogés, alors que dans chaque UPE des secteurs non urbains, 50 ménages étaient interrogés en vue d'une sélection opérée par échantillonnage systématique. Les entretiens de la première phase ont été conduits dans 26 081 ménages dans l'ensemble du pays, où des informations ont été recueillies sur tous les enfants de 5 à 17 ans inclus.

Pour la phase 2, un sous-échantillon systématique d'un maximum de 5 ménages dans les strates urbaines et de 10 ménages dans les strates rurales a été tiré au sort parmi les ménages où il apparaissait à l'évidence qu'au moins un enfant par ménage était engagé dans le travail des enfants. Cela a automatiquement confiné l'échantillon aux UPE dans lesquelles il y avait au moins un enfant travailleur. Les entretiens de la seconde phase ont été conduits dans 4 494 des ménages «éligibles». Au cours de cette phase ont été recueillies des informations sur quelques 10 000 enfants âgés de 5 à 17 ans.

Tri des ménages pour le questionnaire de la phase 2

Comme indiqué ci-dessus, seuls les ménages ayant au moins un enfant engagé dans des activités en relation avec le travail des enfants pouvaient être sélectionnés pour les entretiens complémentaires. Au cours de la phase 1 ont été identifiés les ménages ayant des enfants de 5 à 17 ans dont un était engagé dans un type quelconque de travail. Après l'entretien, l'enquêteur devait attribuer un code d'activité à chaque ménage:

Code d'activité 1: Ménages ayant des enfants engagés dans toute activité effectuée en échange *d'un salaire, d'un profit ou d'un gain économique familial.*

Code d'activité 2: Ménages ayant un enfant engagé seulement dans le transport d'eau et/ou le ramassage de bois, et/ou dans les tâches ménagères et/ou dans l'aide apportée à l'école pendant plus d'une heure par jour.

Code d'activité 3: Ménages n'ayant pas d'enfant engagé dans une activité en relation avec le travail ou ménages sans enfants.

L'objectif majeur de l'enquête était d'examiner les caractéristiques des ménages ayant des enfants engagés dans des activités effectuées en échange d'un salaire, d'un profit ou d'un gain économique familial. Le deuxième objectif majeur était d'examiner les caractéristiques des ménages ayant des enfants engagés seulement dans les tâches ménagères, le travail à l'école, ou le transport d'eau et/ou le ramassage de bois. Dans la seconde phase, notre intérêt majeur était donc d'interroger les ménages ayant reçu le code d'activité 1, alors que notre second intérêt allait à ceux qui avaient reçu le code d'activité 2. Si un ménage avait un ou des enfants qui exerçaient des activités au titre du code 1 comme du code 2, la préférence allait au code 1.

Commentaires

1. Dans cette enquête, l'échantillonnage de la phase 1 à la phase 2 se situe au niveau des ménages. Dans chaque secteur de l'échantillon, la taille d'échantillon de la phase 2 atteint un maximum de 20 pour cent de celle de la phase 1, si ce n'est que l'échantillon de la phase 2 inclut seulement les ménages ayant au moins un enfant qui travaille, alors que la phase 1 couvre tous les ménages comptant plusieurs personnes. Il s'ensuit que tous les secteurs de l'échantillon de la phase 1 sont conservés dans l'échantillon de la phase 2, à l'exception de ceux où il n'y a pas d'enfants qui travaillent.

2. Une autre stratégie d'échantillonnage pourrait consister à réduire le nombre des secteurs de l'échantillon de la phase 2 en introduisant un sous-échantillonnage de la phase 1 à la phase 2 *au niveau du secteur.* (Si besoin est, le taux de sous-échantillonnage des ménages dans les secteurs de l'échantillon peut être accru

proportionnellement pour conserver la taille d'échantillon des ménages souhaitée pour la phase 2.) Ce modèle de substitution peut avoir l'avantage d'abaisser les coûts et d'améliorer le contrôle de qualité, puisque l'échantillon qui en résultera sera moins éparpillé. De plus, avec un système d'échantillonnage approprié, le nombre des secteurs comptant très peu de ménages intéressant l'enquête peut être réduit dans l'échantillon de la phase 2. Bien entendu, la concentration de l'échantillon dans un nombre de secteurs plus limité augmente la variance de l'échantillonnage et les effets du plan de sondage, mais les solutions de substitution mentionnées ci-dessus amélioreront probablement la rentabilité de l'échantillonage.

Sélection des ménages pour le questionnaire de la seconde phase

Les quotas prévus pour les entretiens de la seconde phase étaient de 5 entretiens dans les secteurs urbains et de 10 entretiens dans les secteurs ruraux.

Les 5 ménages des secteurs urbains ont été choisis pour la seconde phase comme suit:

- S'il y avait 5 ménages ou moins ayant les codes d'activité 1 ou 2, ils se retrouvaient tous dans l'échantillon du second stade et aucun échantillonnage n'était nécessaire.
- S'il y avait plus de 5 ménages ayant les codes d'activité 1 ou 2, mais 3 ou moins avec le code d'activité 1, tous les ménages ayant le code d'activité 1 étaient inclus et les ménages ayant le code d'activité 2 étaient échantillonnés pour amener le total à 5 ménages.
- S'il y avait plus de 5 ménages ayant les codes d'activité 1 ou 2, mais 2 ou moins avec le code d'activité 2, tous les ménages ayant le code d'activité 2 étaient inclus et les ménages ayant le code d'activité 1 étaient échantillonnés pour amener le total à 5 ménages.
- Dans tous les autres cas, on a échantillonné les ménages ayant les codes d'activité 1 pour obtenir un échantillon de 3 ménages et ceux ayant le code d'activité 2 pour obtenir un échantillon de 2 ménages, soit un total de 5 ménages.

Pour la seconde phase, les 10 ménages des secteurs ruraux ont été sélectionnés de manière identique, si ce n'est que le nombre des ménages concernés a été doublé partout.

Commentaires

1. L'objectif du système d'échantillonnage ci-dessus est de contrôler la composition de l'échantillon obtenu dans chaque secteur en fonction du code d'activité des ménages.

2. Les procédures d'échantillonnage des ménages sont assez complexes et reposent sur une série d'informations et sur la classification détaillée des listes de ménages. Souvent, une telle rigidité n'est ni nécessaire ni utile, et elle est toujours embarrassante. Avec une certaine flexibilité dans la taille et la composition de l'échantillon obtenu à partir des différents secteurs, on peut concevoir des systèmes de substitution permettant d'obtenir des échantillons probabilistes assurant en même temps un très bon contrôle de la taille et de la composition de l'échantillon

tout entier. Cela devient plus facile à réaliser quand il y a aussi un sous-échantillonnage de la phase 1 à la phase 2 au niveau du secteur.

3. Le poids des échantillons doit être calculé pour les ménages de chaque catégorie et dans chaque secteur. Cela exige un dénominateur du taux d'échantillonnage - le nombre réel de ménages (à l'exclusion des espaces vides) dans cette catégorie de la liste -, ce qui implique des contraintes en matière d'informations nécessaires et des risques d'erreur, et demande dans tous les cas beaucoup de temps. Il existe des méthodes plus simples et meilleures.

2.9. Enquête sur l'activité des jeunes en Jamaïque 2002: description et commentaires

Les informations suivantes ont été extraites du rapport de l'Enquête sur l'activité des jeunes en Jamaïque 2002 menée par l'Institut de statistique de la Jamaïque. C'est là un exemple d'enquête modulaire dont la taille, la conception et le calendrier ont été entièrement déterminés par ceux de l'enquête auprès des parents à laquelle elle était liée.

2.9.1. Objectifs et contenu

Objectifs de l'enquête

L'Enquête sur l'activité des jeunes en Jamaïque (YAS) a été conduite sur le plan national en tant que *module* de l'Enquête sur la main-d'œuvre d'avril 2002. Dans tous les ménages interrogés pour l'enquête sur la main-d'œuvre, lorsque l'on a constaté que des enfants du groupe d'âge de 5 à 17 ans vivaient dans le ménage, ces enfants remplissaient les conditions requises pour être interrogés dans le cadre de l'enquête sur l'activité des jeunes.

L'objectif général de la YAS était de déterminer le caractère, la nature, la dimension et les raisons du travail des enfants en Jamaïque, ainsi que les conditions de travail et leurs effets sur la santé, l'éducation et le développement normal de l'enfant travailleur. Plus spécifiquement, l'enquête a permis d'obtenir des informations concernant:

1. Les caractéristiques démographiques et socio-économiques, telles que le niveau d'instruction et de formation (inscription et fréquentation), la profession et les compétences, les horaires de travail, la rémunération et autres conditions de travail et de vie.

2. Les caractéristiques des secteurs où travaillent des enfants: secteur public/privé.

3. Où travaillent les enfants, et depuis combien de temps; quels sont les facteurs qui incitent les enfants à travailler et/ou les familles à les mettre au travail.

4. Les perceptions des parents/tuteurs, des enfants et des employeurs au sujet du travail des enfants, de la réglementation, des lois et de la législation, etc.

5. La participation à des programmes ayant un impact positif sur l'élimination du travail des enfants.

6. La situation des enfants travailleurs en matière de santé, de protection sociale et d'éducation.

Population cible

La population cible de l'enquête était la population des enfants du groupe d'âge de 5 à 17 ans vivant dans des ménages privés. Les enfants vivant dans des institutions comme les foyers, les hôpitaux et les centres d'accueil n'ont pas été couverts par l'enquête. Ce groupe d'âge cible a été choisi en fonction des considérations suivantes:

- Borne supérieure: la Convention des Nations Unies relative aux droits de l'enfant (1989) et la Convention de l'OIT sur les pires formes de travail des enfants (1999) ont défini les enfants comme étant les personnes âgée de moins de 18 ans.
- Borne inférieure: il était nécessaire d'obtenir des informations sur les enfants n'ayant pas l'âge de la scolarité obligatoire, mais la limite de 5 ans a été choisie parce qu'il ne pouvait y avoir que très peu d'enfants travailleurs en dessous de cet âge.

Commentaire

Selon les objectifs énoncés, en dépit du champ considérable des informations recherchées, l'enquête est clairement centrée sur le travail des enfants.

2.9.2. L'entretien d'enquête

Les informations recueillies par l'Enquête sur l'activité des jeunes ont été obtenues à l'occasion d'entretiens avec les adultes et avec les enfants concernés. Dans le cas de l'adulte, le répondant de référence était le parent/tuteur de l'enfant. Les adultes ont été interrogés sur les activités économiques et non économiques de chaque enfant, ainsi que sur le ménage en général. Les enfants ont été interrogés sur leurs activités économiques et non économiques. Les enquêteurs devaient d'abord remplir les questionnaires de l'enquête sur la main-d'œuvre destinés aux ménages, puis le questionnaire de l'enquête sur l'activité des jeunes destiné aux enfants concernés.

Ce questionnaire a été spécifiquement conçu pour recueillir des informations détaillées sur les enfants âgés de 5 à 17 ans inclus. Des informations ont été collectées sur l'éducation des enfants, leurs activités économiques et non économiques présentes/passées, leur profession, leur branche d'activité, leurs revenus, leurs horaires de travail, etc. Le questionnaire était divisé en 9 sections:

Section 1: Fichier ménages ayant des enfants de 5 à 17 ans

Section 2: Éducation et formation professionnelle

Section 3: 3A – Activité économique actuelle et Section 3B – Activité économique habituelle

Section 4: Activité non économique actuelle

Section 5: Questions de santé et de sécurité en relation avec le travail

Section 6: Perception de l'enfant par les parents

Section 7: Statut migratoire de l'enfant

Section 8: Caractéristiques du ménage

Section 9: Questions à adresser aux enfants de 5 à 17 ans.

Les réponses aux sections 1-8 devaient être apportées par le parent/tuteur de l'enfant ciblé: si cette personne n'était pas disponible, on choisissait un autre adulte responsable de plus de 17 ans. Les questions de la section 9 devaient être adressées à chaque enfant du groupe d'âge de 5 à 17 ans.

La première section du questionnaire à remplir par l'enquêteur était la section 1 – le fichier des ménages, qui listait les noms et les âges de tous les enfants concernés dans le ménage. Il devait être copié par l'enquêteur dans les informations fournies par l'enquête sur la main-d'œuvre. Tous les enfants ont été listés sur le fichier avant le début des entretiens, ce qui assurait qu'aucun enfant concerné ne soit exclu de l'enquête.

Pour chaque enfant concerné, les entretiens avec l'adulte ont été répétés pour les sections 1-7. Une fois ces entretiens achevés, les informations sur le logement – section 8 – ont été collectées auprès des adultes.

Les enfants concernés ont ensuite été interrogés, et chacun s'est vu poser les questions de la section 9. Lorsque c'était possible, les enfants ont été interrogés par ordre d'âge, selon l'ordre du listing sur le fichier des ménages. Le questionnaire a donc suscité des réponses provenant à la fois des parents/tuteurs et des enfants eux-mêmes. Toutefois, il y a eu d'importantes discordances entre les uns et les autres quand les mêmes questions leur étaient posées, ce qui peut en partie être dû à l'âge des enfants. Le rapport d'enquête a donc utilisé essentiellement les questions adressées au parent/tuteur, excepté dans la situation où la question était seulement posée à l'enfant.

Commentaire

On notera qu'à part les questions spécifiquement liées au travail des enfants, ce module est appliqué à tous les enfants du groupe d'âge spécifié. Il peut donc être décrit de la manière la plus appropriée comme un module «d'étude des enfants» plutôt que comme un module «d'étude du travail des enfants». Cet arrangement présente l'avantage de collecter des informations comparables sur tous les enfants, si bien que ceux qui travaillent peuvent être comparés à ceux qui ne travaillent pas afin de mieux identifier les conditions et les conséquences du travail des enfants. Également important est le fait qu'il permette l'exploration de toute la série des activités des enfants, et non pas simplement de ce qui est qualifié d'«activité économique». L'engagement substantiel des enfants dans les travaux ménagers revêt une importance particulière.

L'inconvénient majeur de cette approche est qu'une part d'efforts disproportionnée est consacrée à enquêter sur les enfants non travailleurs, aux dépens des enfants travailleurs qui sont le groupe cible essentiel. Cela est particulièrement vrai quand seule une petite proportion des enfants est engagée dans le travail des enfants.

Ces informations peuvent être extrêmement précieuses en elles-mêmes: les commentaires ci-dessus s'appliquent à la pertinence des données pour l'étude du travail des enfants.

2.9.3. Aspects de l'échantillonnage

La conception de l'échantillon des Enquêtes sur la main-d'œuvre en Jamaïque

Les Enquêtes sur la main-d'œuvre de la Jamaïque sont partie intégrante du programme continu d'enquêtes sur les ménages de l'Institut de statistique de la Jamaïque: elles sont effectuées régulièrement depuis les années 1960. Ces enquêtes ont lieu chaque année à échéance trimestrielle - en janvier, avril, juillet et octobre -, la semaine de référence habituelle étant l'ensemble de la dernière semaine de travail du trimestre précédent. Elles portent sur un échantillon de 1 % des ménages privés, qui sont choisis selon un modèle aléatoire stratifié. Sont exclus de l'échantillon les ménages non privés qui incluent les logements collectifs, tels que les camps militaires, les pensionnats, les cliniques psychiatriques, les hôpitaux et les prisons.

Dans l'Enquête sur la main-d'œuvre de la Jamaïque, la population active est constituée par les personnes âgées de 14 ans et plus qui travaillent, cherchent du travail ou souhaiteraient travailler, et qui ne sont pas scolarisées à plein temps. Les entretiens avec les enfants de moins de 14 ans sont limités à la collecte de données sur l'âge, le sexe et la parenté avec le chef du ménage.

Pour l'Enquête sur l'activité des jeunes 2002, le plan d'échantillonnage a été basé sur celui de l'Enquête sur la main-d'œuvre 2002. Le plan de base de l'enquête était celui d'un échantillonnage systématique stratifié à deux niveaux, le premier niveau étant une sélection de secteurs dits «de dénombrement» (SD), et le second une sélection d'unités d'habitation au sein des SD sélectionnés.

Premier niveau de la sélection

La conception de l'EMO était essentiellement centrée sur la sélection d'un échantillon représentatif distribué de manière proportionnelle entre les secteurs urbains et ruraux sur la base du nombre de secteurs de dénombrement. La situation idéale aurait été de sélectionner l'échantillon sur une base proportionnelle à la taille de la population, mais cette information n'était pas disponible. Les données démographiques du recensement de 1991 étaient devenues obsolètes à cause du nombre de changements démographiques intervenus dans le pays, et celles du recensement de 2001 n'étaient pas encore accessibles.

La base de sondage utilisée pour la sélection du premier niveau est une liste de SD stratifiés en SD urbains et ruraux. Cette liste était parfaitement à jour, car elle avait été modifiée avant le recensement de 2001. Pour la sélection du deuxième niveau, un listage des SD sélectionnés a été effectué avant l'enquête. Pour cet exercice, chaque unité d'habitation des SD sélectionnés a été visitée par des enquêteurs, et des informations appropriées relatives à la composition et à la description de l'unité d'habitation ont été collectées. … Les SD sont stratifiés en 2 543 SD urbains et 2 693 SD ruraux. Chaque sélection a été faite sur la base de la commune sans aucun recoupement et définie de manière à inclure l'île tout entière. Les SD ont ensuite été

réunis en unités primaires d'échantillonnage (UPE), définies comme comportant un minimum de 80 habitations. Le second niveau a consisté à sélectionner des habitations en utilisant à nouveau une méthode d'échantillonnage systématique à partir d'une liste actuelle des habitations à l'intérieur de l'UPE.

Le processus de sélection du premier niveau a adopté une méthode d'attribution proportionnelle dans laquelle un échantillon de 10 pour cent des SD a été sélectionné pour chaque commune et attribué proportionnellement aux strates urbaines/rurales. Cette sélection est basée sur la procédure suivante:

1. Les secteurs de dénombrement sont groupés en strates urbaines et rurales dans chaque commune.

2. Le nombre des SD à sélectionner dans les strates urbaines et rurales de chaque commune est calculé comme suit:

$$m_{hu} = m_h . \left(\frac{M_{hu}}{M_h} \right) = m. \left(\frac{M_{hu}}{M} \right),$$

$$m_{hr} = m_h . \left(\frac{M_{hr}}{M_h} \right) = m. \left(\frac{M_{hr}}{M} \right),$$

dans lesquelles:

m_{hu} = le nombre de SD à sélectionner sur un total de M_{hu} de la strate urbaine (u) de ménages de la commune *h*

m_{hr} = le nombre de SD à sélectionner sur un total de M_{hr} de la strate rurale (r) de ménages de la commune *h*

$M_h = M_{hu} + M_{hr}$, $m_h = m_{hu} + m_{hr}$, le total des SD et le nombre total sélectionné dans les ménages de la commune *h*

$M = \Sigma_h M_h$, $m = \Sigma_h m_h$, le total des SD et le nombre total sélectionné dans l'ensemble du pays.

3. Les SD des strates urbaines et rurales de chaque commune sont sélectionnés selon une procédure systématique.

Second niveau de la sélection

Au second niveau de l'échantillonnage, on a procédé à un listage de toutes les habitations des SD sélectionnés au premier stade. La liste a été rendue circulaire et un échantillon de 32 habitations a été sélectionné systématiquement. Après la sélection, ils ont été placés dans des panels de taille égale dont 16 ont été utilisés pour l'enquête.

Commentaires

1. On notera que cette procédure n'aboutit pas à un échantillon autopondéré de ménages ou de personnes. Si les SD sont sélectionnés avec une probabilité constante dans tout le pays, les ménages au sein d'un SD sont sélectionnés avec

une probabilité inversement proportionnelle au nombre total de ménages du SD. Les ménages des SD les plus importants ont donc moins de chances de sélection que les ménages des SD plus restreints. La différence ne sera faible que si les SD sont de taille uniforme.

2. On notera aussi qu'aux différences de taille moyenne des SD entre domaines (entre secteurs urbains et ruraux) correspondront des différences semblables dans la répartition et les pourcentages de l'échantillonnage.

3. La pondération de l'échantillon doit être calculée d'après le listage. Le nombre des ménages listés, à l'exclusion des espaces blancs, doit être connu et enregistré à cette fin.

Enquête modulaire sur le travail des enfants

L'échantillon de la YAS a été entièrement déterminé par celui de l'EMO à laquelle elle était rattachée en tant que module. Mis à part l'adjonction de ce module au questionnaire d'EMO, seules les modifications suivantes ont été apportées aux procédures indiquées pour l'EMO:

1. L'Enquête sur l'activité des jeunes 2002 était attachée à la phase d'avril de l'Enquête sur la main-d'œuvre 2002. Du fait que l'Enquête sur l'activité des jeunes était un module de l'Enquête sur la main-d'œuvre, la période de référence a été modifiée pour accueillir l'Enquête sur l'activité des jeunes. La semaine de référence utilisée pour les deux enquêtes allait du dimanche 17 mars au samedi 23 mars 2002 et non pas, selon l'habitude, du 24 au 30 mars 2002. Ce changement a été effectué parce que les enfants auraient été en vacances scolaires de Pâques pendant la période de référence habituelle, ce qui aurait eu des incidences sur les réponses aux questions concernant la fréquentation scolaire pendant la période de référence.

2. Des dispositions spéciales ont été prises pour contrôler la distribution de la population (groupe d'âge de 5 à 17 ans) dans la post-stratification de la pondération de l'échantillon, comme il est indiqué ci-dessous (section 2.9.5).

Commentaire

Une question très importante est de savoir si - et éventuellement dans quelle mesure – la taille, la conception et les opérations de l'enquête auprès des parents sont modifiées pour être mieux appropriées aux conditions du module sur le travail des enfants. L'ampleur des modifications possibles à cet effet est plus limitée quand l'enquête auprès des parents, dans le cas présent la phase d'avril de l'EMO 2002, constitue un élément d'une série régulièrement échelonnée dans le temps.

Même quand d'autres aspects de la conception demeurent inchangés, l'adjonction du module sur le travail des enfants augmente la durée de l'entretien, ce qui peut influer sur la qualité des données collectées dans le cadre de l'enquête sur les parents. Cela limite le champ de l'information additionnelle qui peut être collectée dans le module sur le travail des enfants.

Pondération de l'échantillon

Le but essentiel de la pondération est de procéder à des ajustements en fonction des probabilités différentielles de sélection utilisées dans le processus d'échantillonnage pour réduire la variabilité de l'échantillonnage et compenser tout sous-dénombrement ou non-réponse dans l'enquête.

Les habitations échantillonnées qui étaient dans le champ de l'enquête, où aucune collecte d'information n'était possible parce qu'elles étaient vides ou fermées ou que leurs occupants refusaient de coopérer au moment de l'enquête, ont été classées parmi les non-répondants. Le but de cet ajustement était de réduire la distorsion résultant du fait que les non-répondants sont différents de ceux qui répondent à l'enquête.

La post-stratification est une méthode utilisée pour ajuster les estimations pondérées de l'enquête afin de les faire concorder avec les estimations de la population. Elle est généralement utilisée pour compenser les taux de réponse différents des sous-groupes démographiques, augmenter la précision des estimations de l'enquête et réduire les distorsions qu'elles comportent. La méthode utilisée a été celle d'un simple ajustement de ratios qui a été appliqué à la pondération de l'échantillonnage en utilisant la projection de la population totale 2002 pour chaque strate définie par âge et par sexe et pour chaque commune.

La méthode de pondération utilisée pour la YAS a assuré la conformité des estimations avec la distribution de la population (groupe d'âge de 5 à 17 ans) par âge, par sexe et par commune.

Commentaires

Il est erroné de classer parmi les non-répondants les habitations vides et inoccupées: il s'agit d'espaces blancs qui ne doivent pas être pris en compte, très différents des refus et autre cas de non-réponse authentique, pour lesquels une compensation par pondération est appropriée.

Taille de l'échantillon

La taille d'échantillon de la YAS a été déterminée par celle de l'EMO.

Le rapport d'enquête indique que l'échantillon était trop petit pour faciliter une analyse statistique inférentielle. Si les enquêtes régulières sur la main-d'œuvre contiennent quelques informations sur le travail des enfants de 14 ans et plus, il existe peu d'informations concernant les groupes d'âge plus jeunes. En 1994, une étude préliminaire menée en Jamaïque a permis d'estimer à 23 000 le nombre des enfants de 6 à 16 ans qui étaient engagés dans des activités de travail des enfants. Pour un échantillon à 1 pour cent, cela devrait représenter environ 230 cas enquêtés. En fait, dans l'échantillon final de la YAS, qui comptait 6 189 enfants, seuls 143 ont indiqué avoir travaillé pendant la semaine de référence. L'échantillon d'enfants travailleurs était donc restreint et a limité la série des analyses.

De plus, il convient d'interpréter avec prudence les résultats concernant la participation des enfants à des activités économiques. Comme nous l'avons indiqué ci-dessus, si le questionnaire a fourni des réponses à la fois des parents/tuteurs et des enfants

eux-mêmes, il y a eu d'importantes discordances entre les uns et les autres quand des questions similaires leur étaient posées, ce qui peut être dû en partie à l'âge des enfants.

Commentaires

1. La proportion d'enfants engagés dans le travail est apparue en fait, dans l'enquête principale, beaucoup plus faible qu'il n'avait été rapporté dans le pré-test. On comparera les chiffres mentionnés ci-dessus pour l'enquête principale avec ceux du pré-test, tels que les présente le rapport d'enquête:

   > «Dans le pré-test, il y avait un total de 340 enfants provenant de 162 ménages interrogés dans les quatre communes de Kingston, St. Andrew, St. Elizabeth et St. Catherine. … Environ 10,5 pour cent des enfants de l'échantillon de pré-test avaient travaillé au cours des 12 derniers mois. Sur les 36 enfants qui avaient travaillé, 22 avaient été employés. Le nombre des enfants qui avaient travaillé au cours de la semaine précédente était de 17, soit 5 pour cent de l'échantillon. Parmi ceux-ci, 11 étaient employés dans le commerce de gros et l'agriculture commerciale. La majorité des enfants (75,7 pour cent) aidaient aux tâches domestiques dans les ménages du pré-test, bien qu'ils aient été un pourcentage notable (23,1 pour cent) à indiquer qu'ils ne le faisaient pas.»

   Mis à part la variabilité d'échantillonnage que traduisent ces chiffres, cette différence entre les résultats du pré-test et ceux de l'enquête principale peut être due en partie au fait que l'échantillon du pré-test n'était pas représentatif du pays tout entier. Toutefois, elle peut aussi provenir du risque plus considérable d'omission inhérent à une opération d'une envergure beaucoup plus importante.

1. Si la taille d'échantillon disponible pour le module d'étude du travail des enfants est déterminée par la taille d'échantillon de l'enquête auprès des parents, la situation peut être améliorée, quand – comme dans le cas présent – cette dernière est une enquête régulière, en répétant ce module dans plus d'une phase de l'enquête.

2. Il peut également être utile de prendre en compte dans l'analyse une plus large définition du travail des enfants: d'inclure non seulement les enfants engagés dans une activité économique au sens strict, mais aussi ceux qui apportent une aide «substantielle» définie de manière appropriée dans les tâches ménagères.

Chapitre 3
L'échantillonnage en vue d'une enquête typique basée sur la population

3.1. Introduction

La *collecte d'informations sur le travail des enfants en conjonction avec une vaste enquête sur la population générale* telle que l'enquête sur la main-d'œuvre (EMO), revêt très souvent la forme de questions sur le travail des enfants attachées à l'enquête de base en tant que module. La collecte des deux types de données peut aussi comporter des dénombrements séparés, mais basés sur un modèle commun, l'enquête sur le travail des enfants (ETE) formant généralement un sous-échantillon de l'enquête de base.

Si l'objectif ici n'est pas d'examiner les principes généraux de la conception d'un échantillon applicables à une EMO ou à une ETE similaire, ce chapitre a pour but de *clarifier les procédures d'échantillonnage que comporte le type le plus courant de modèles utilisé pour ces enquêtes*. Cet aspect est important lorsque l'on cherche à comprendre comment les différents systèmes d'estimation du travail des enfants peuvent être liés à ce modèle ou en être dérivés.

Les sections 3.2 à 3.4 abordent certains aspects pratiques importants du plan d'échantillonnage et de la base de sondage, parmi lesquels la stratification et l'échantillonnage à degrés multiples. La question importante mais complexe du choix de la taille d'échantillon est examinée assez longuement dans la section 3.5. Les sections 3.6 à 3.10 sont consacrées aux procédures pratiques de tirage de l'échantillon, incluant l'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT) en présence de dimensions imparfaites de la taille, et traitant des unités primaires d'échantillonnage (UPE) très grandes et très petites. Enfin, la section 3.11 fournit une illustration chiffrée des procédures, sur la base de données simulées.

3.2. Le plan d'échantillonnage: orientation pratique

L'objectif de ce chapitre est de mettre en lumière quelques bonnes pratiques d'échantillonnage dans la conception et la réalisation des enquêtes sur le travail des enfants. Nous devons partir de certains principes de base. La condition importante est ici, au moins en ce qui concerne les enquêtes standard auprès des ménages sur le travail des enfants, qu'elles soient basées sur des échantillons «probabilistes et mesurables» tels qu'ils sont décrits ci-dessous.

3.2.1. Les raisons de l'enquête par échantillonnage

Les enquêtes par échantillonnage sont l'une des principales sources d'information statistique. Leur importance est encore accrue en cas de défaut d'autres sources d'information, comme les dossiers et registres administratifs dans les pays en développement. Ce type d'enquête implique que l'information soit obtenue seulement sur un sous-ensemble d'unités de la population duquel on puisse tirer des inférences concernant la population tout entière. Il est clair que la validité de ces inférences dépendra de la manière dont sera prélevé l'échantillon, ainsi que de la taille de ce dernier.

Comparée à une couverture complète ou par recensement, la méthode de l'échantillonnage peut présenter de nombreux avantages: ceux-ci résultent du fait que la restriction de la collecte d'information à un échantillon réduit l'échelle de l'opération. Par conséquent, les coûts sont également réduits et les résultats peuvent être obtenus plus rapidement; il est possible de rechercher des informations plus élaborées, et d'employer des méthodes plus intensives qui fourniront des données plus exactes. Un autre avantage majeur des enquêtes par échantillonnage réside dans leur souplesse beaucoup plus grande par comparaison avec les recensements complets: souplesse en termes de calendrier, de couverture, de taille, de contenu et autres aspects de la conception et de la méthodologie.

La principale limite de la méthode par échantillonnage réside dans la variabilité d'échantillonnage à laquelle sont sujettes les estimations de l'enquête. Par-dessus tout, ceci limite la mesure dans laquelle les résultats de l'enquête peuvent être décomposés géographiquement, par sous-groupe de population ou dans le temps, à cause de la variabilité d'échantillonnage résultant d'une taille limitée. Cela est illustré de manière particulièrement claire par l'impossibilité de la plupart des enquêtes par échantillonnage à produire des estimations directes pour les zones locales et autres domaines restreints.

Il est également important de souligner que ces avantages et ces limites bien connus de la méthode de l'échantillonnage par rapport au recensement s'appliquent également lorsque l'on considère des échantillons de tailles différentes. Les enquêtes plus restreintes peuvent être moins coûteuses, plus opportunes, plus intensives en procédures, plus riches en contenu, aptes à fournir des estimations plus précises, et susceptibles d'être répétées dans le temps. Par contre, elles sont sujettes à des erreurs d'échantillonnage plus importantes, qui limitent les possibilités de s'appuyer sur les estimations produites, et notamment de décomposer les résultats de ces enquêtes sur le plan géographique, par sous-groupe de population ou dans le temps. Nous examinerons cette question fondamentale du choix de la taille d'échantillon dans la suite de ce chapitre.

La conception des échantillons est une tâche hautement technique. Elle implique dans ses grandes lignes de déterminer:

1. La taille d'échantillon, c'est-à-dire le nombre d'unités d'analyse à sélectionner pour l'enquête;

2. la structure de l'échantillon, c'est-à-dire le mode de sélection de ces unités;

3. les procédures d'estimation, c'est-à-dire la manière dont les résultats doivent être utilisés pour tirer des inférences sur toute la population concernée dans laquelle l'échantillon a été choisi.

La conception de l'échantillon ne peut pas être isolée des autres aspects de la conception et de la réalisation de l'enquête. Dans la pratique, la conception de l'échantillon doit prendre en compte de nombreuses considérations autres que la théorie de l'échantillonnage. *La taille et la structure de l'échantillon déterminent l'ampleur non seulement de l'erreur d'échantillonnage, mais aussi de la plupart des autres composantes de l'erreur; bien entendu, elles déterminent aussi le coût de l'enquête.* Par exemple, si l'on a essayé de minimiser l'erreur d'échantillonnage en menant une enquête d'entretiens personnels sur un échantillon aléatoire de personnes à travers le pays, les coûts des déplacements peuvent apparaître comme prohibitifs et la supervision du travail des enquêteurs comme difficile. De plus, selon la situation, il peut arriver que l'on ne dispose pas de listes exactes et complètes des personnes pour la sélection de l'échantillon, et il est courant que l'on ne parvienne pas à localiser les individus choisis sur le terrain. Tous ces facteurs ont souvent pour résultat une importante augmentation des coûts et des erreurs.

Il est vrai que les composantes de l'erreur d'enquête qui ne relèvent pas de l'échantillonnage dépendent aussi de nombreux facteurs autres que la taille et la structure de l'échantillon. C'est toutefois une erreur – malheureusement répandue – que de supposer l'impact primaire de la taille et de la structure de l'échantillon limité à l'ampleur de l'erreur d'échantillonnage. Il est clair qu'au-delà d'une certaine limite, *l'augmentation de la taille d'échantillon peut avoir un effet négatif sur tous les aspects de la qualité des données:* pertinence de l'enquête (en restreignant la somme d'information qui peut être collectée de manière gérable pour chaque cas dans un échantillon de grande dimension); opportunité (délai séparant la collecte et le traitement d'un volume d'information accru); et bien entendu qualité de sa réponse (formation, supervision et contrôle inadéquats de l'enquêteur, dus à l'augmentation du volume de travail sur le terrain). Une structure d'enquête inappropriée (par exemple, un éparpillement géographique excessif de l'échantillon, la répétition des entretiens et du suivi des enquêtés, etc.) peut avoir un effet semblable sur la qualité des données.

Ces considérations imposent la nécessité d'une orientation absolument pratique de la conception des échantillons. En premier lieu, il est souhaitable de dépendre de méthodes et de procédures qui soient fiables et pratiques dans ces circonstances particulières, afin de faire en sorte que les performances réelles sur le terrain et dans les bureaux ne s'écartent pas trop des conditions de la conception. Une longue chaîne d'opérations va de la conception à la sélection d'échantillons, à la collecte de données sur le terrain, puis elle revient au codage, aux calculs et à l'analyse statistique. Pour achever ce circuit avec un degré de sécurité raisonnable, *les procédures d'échantillonnage doivent être robustes, c'est-à-dire insensibles au moindre écart modeste par rapport à l'idéal.* Il est préférable d'adopter une conception et des procédures qui, même si elles ne sont pas les plus précises et les plus raffinées dans des conditions optimales, peuvent néanmoins permettre de faire face aux situations inattendues et inconnues que l'on rencontre invariablement dans le travail d'enquête pratique. L'échantillonnage pratique doit être inscrit dans la nature et les interrelations

des opérations de terrain et administratives non seulement dans la mise en œuvre d'un certain modèle d'échantillon, mais aussi dans la totalité du stade de la collecte de données. L'échantillonneur doit constamment rester conscient de ce qui peut et ne peut pas être attendu des agents d'exécution et des administratifs qui participent réellement à la mise en place de l'échantillon.

La nécessité d'une orientation pratique ne signifie pas que la théorie de l'échantillonnage soit sans importance. Au contraire, on ne peut être un bon échantillonneur pratique sans avoir une bonne compréhension de la théorie sous-jacente qui guide le travail pratique.

3.2.2. Échantillons probabilistes et mesurables

Sur une base scientifique, on ne peut déduire des inférences de l'échantillon à la population tout entière que si l'échantillon est composé d'unités sélectionnées selon une procédure aléatoire qui donne une chance de sélection connue et différente de zéro à chaque unité de la population cible. Un échantillon établi de cette manière est appelé *échantillon probabiliste*. L'expression *échantillon aléatoire* est couramment utilisée pour désigner la même réalité.

La conception d'un échantillon probabiliste spécifie le type de procédure aléatoire appliquée dans la sélection de l'échantillon. Elle spécifie aussi la manière dont doivent être estimés les paramètres de population à partir des résultats de l'échantillon. Les procédures de sélection et d'estimation constituent deux aspects du *plan de sondage.*

Concernant la procédure de sélection, de nombreux types de plan sont possibles et utilisés dans la pratique. La procédure peut par exemple donner les mêmes chances (égales) d'apparaître dans l'échantillon à tous les éléments de la population, ou bien certaines unités peuvent avoir des chances plus importantes que les autres. Nous pouvons sélectionner les éléments individuellement, ou les grouper d'abord en *grappes* plus importantes et appliquer la procédure de sélection à ces grappes. Nous pouvons partager la population en *strates* et appliquer séparément l'une des procédures quelconque ci-dessus au sein de chaque strate. Chaque procédure aléatoire détermine en fait un ensemble différent d'unités qui peuvent en principe être sélectionnées en utilisant cette procédure, ainsi que la chance de sélectionner un échantillon particulier parmi elles. Cependant, pour être aléatoire, toute procédure de sélection doit veiller à ce que chaque unité de la population reçoive une chance non nulle d'apparaître dans l'échantillon à sélectionner.

La procédure d'estimation fait intervenir des formules statistiques ou mathématiques sur la base de valeurs d'échantillon, et éventuellement aussi d'informations provenant d'autres sources extérieures à l'échantillon; elle fournit des estimateurs qui sont utilisés pour produire des estimations des paramètres de population d'intérêt. Cette procédure inclut aussi l'estimation des mesures d'incertitude («erreurs d'échantillonnage», «intervalles de confiance», etc.) auxquelles sont sujets les résultats des sondages.

3.2.3. Obtenir un échantillon probabiliste

Comme on l'a indiqué, toute enquête visant à mettre en pratique à toute la population d'intérêt des observations effectuées sur un échantillon doit être basée sur l'échantillonnage probabiliste. Pour obtenir un échantillon aléatoire, il convient de suivre certaines procédures appropriées aux stades de la sélection, de l'exécution et de l'estimation. Ces procédures sont les suivantes:

1. Chaque élément de la population doit être représenté explicitement ou implicitement dans la base à partir de laquelle l'échantillon est sélectionné: pour cela, nous avons besoin d'une bonne *base de sondage* (section 3.3).

2. L'échantillon doit être sélectionné dans cette base selon un processus comportant un ou plusieurs degrés de randomisation automatique, qui donne à chaque unité une probabilité de sélection spécifiée. Cela implique de spécifier des procédures de sélection clairement établies avant la sélection effective des unités qui composeront l'échantillon, et de ne pas modifier l'échantillon après sa sélection.

3. Au stade de la mise en œuvre, toutes les unités sélectionnées – et seulement celles-ci – doivent être incluses dans l'enquête et correctement dénombrées. Cela implique d'éviter les non-réponses et de n'autoriser aucune substitution pour les non-répondants.

4. Dans l'estimation des valeurs de population à partir de l'échantillon, les données provenant de chaque unité de l'échantillon devraient être pondérées conformément à la probabilité de sélection de l'unité, ou *des pondérations du plan.* En fait, il est souvent nécessaire de modifier *les pondérations du plan* pour limiter l'impact des autres insuffisances de l'échantillon: la série de pondérations appliquée à chaque unité dénombrée dans l'échantillon permet d'obtenir les estimations correspondantes pour la population. Les procédures de calcul de ces pondérations sont décrites au chapitre 7 ci-dessous.

Toutefois, dans la pratique, certaines approximations dans l'application de ces exigences idéales sont souvent nécessaires pour des raisons comme par exemple:

- non-inclusion de certaines unités dans la base de sondage (sous-couverture);
- distorsions dans les probabilités de sélection dues à d'autres erreurs de couverture et de sélection de l'échantillon;
- impossibilité de dénombrer ou d'obtenir des informations complètes sur toutes les unités sélectionnées (non-réponse); et
- utilisation de procédures approximatives au stade de l'estimation, et en particulier prise en compte incomplète de la méthode de sélection réellement utilisée (biais d'estimation).

Le niveau d'erreur jusqu'auquel un échantillon peut encore être considéré comme un échantillon probabiliste est affaire de jugement pratique.

Le point fort de l'échantillonnage probabiliste réside dans le fait que le mécanisme de sélection aléatoire permet d'appliquer la théorie statistique pour obtenir des estimations

des valeurs de population à partir d'observations de l'échantillon d'une manière essentiellement objective, libre d'hypothèses arbitraires ou de jugements subjectifs.

3.2.4. Les échantillons mesurables

Un concept semblable mais moins exigeant et inclusif est celui de *mesurabilité*. Un échantillon est dit mesurable s'il fournit des estimations non seulement des paramètres de population requis (comme le fait un échantillon probabiliste) mais aussi de leur variance d'échantillonnage. En d'autres termes, un échantillon est mesurable si, à partir de la variabilité observée entre unités au sein de l'échantillon, il est possible d'obtenir des estimations de la variance d'échantillonnage (c'est-à-dire de la variabilité entre les différents échantillons possibles). Pour être mesurable, il est normalement nécessaire que l'échantillon soit un échantillon probabiliste; il devra également remplir certaines autres conditions garantissant que la variance d'échantillonnage peut être estimée d'après la variabilité observée entre unités dans le seul échantillon qui soit disponible. De plus, des hypothèses et des approximations peuvent intervenir dans les procédures d'estimation de variance sans qu'il y ait nécessairement perte de mesurabilité de l'échantillon au sens pratique du terme.

La condition essentielle est que l'échantillon soit conçu, sélectionné et mis en œuvre, et que les informations sur sa structure soient enregistrées, ainsi il peut être fractionné en «reproductions» telles que la variabilité observée entre ces reproductions (ou une fonction de celle-ci) produise une estimation utile dans la pratique («valide») de la variance d'échantillonnage. On trouvera dans l'un des chapitres suivants certaines procédures pratiques d'estimation de la variance.

3.2.5. L'échantillonnage pour une enquête «typique» basée sur la population

Les sections suivantes présentent des descriptions des caractéristiques techniques du plan d'échantillonnage le plus couramment utilisé pour les enquêtes sur la main-d'œuvre (EMO) et autres enquêtes semblables basées sur la population. Les enquêtes sur le travail des enfants peuvent être basées sur le même échantillon, ou sur un échantillon lié à ces enquêtes basées sur la population ou dérivé de celles-ci.

La plupart des échantillons de ménages et de personnes sont choisis au cours d'un certain nombre de stades de l'échantillonnage. Ainsi, dans une enquête nationale auprès des ménages, tout le pays peut être divisé en unités de zone telles que les localités ou les zones de dénombrement (ZD), et un échantillon de ces zones sera sélectionné dans un premier stade. Les types d'unités sélectionnées au cours de ce premier stade sont appelés *unités d'échantillonnage primaires* (UEP). Pour le premier stade de la sélection, il est nécessaire de disposer d'une base d'UEP qui liste les unités couvrant l'ensemble de la population de manière exhaustive et sans chevauchements, et qui fournisse par ailleurs des informations pour une sélection efficiente des unités. Cette base est appelée *base d'échantillonnage primaire* (BEP). Le second stade peut consister à diviser chacune des UEP sélectionnées dans le premier stade en zones plus restreintes comme des îlots, puis à sélectionner une ou plusieurs de ces unités du deuxième stade (UDS) dans chaque UEP sélectionnée. Ce processus peut se poursuivre jusqu'à l'obtention d'un échantillon d'*unités finales de secteur* (UFS) assez restreint. Il

est très fréquent d'utiliser des modèles ne comportant qu'un seul stade sectoriel – ce qui est le cas, par exemple, de la plupart des enquêtes sur le travail des enfants menées avec l'appui du BIT.

Au dernier stade, dans chaque secteur d'échantillon (ou UFS) choisi, les ménages peuvent être listés individuellement et un échantillon peut être choisi avec des ménages ou des personnes comme unités finales d'échantillonnage (UFE). Au cours de l'enquête, l'information peut être collectée et analysée pour les UFE elles-mêmes, ou pour d'autres types d'unités («éléments») associés aux UFE sélectionnées, comme les individus ou les enfants au sein de ménages de l'échantillon.

Le plan d'échantillonnage comporte le choix du nombre de stades à utiliser et, à chaque stade, le type d'unités, la méthode de sélection et le taux d'échantillonnage ou le nombre d'unités à sélectionner. Cela exige une base de sondage qui représente la population cible.

3.3. La population de l'enquête: la base de sondage

3.3.1. La population de l'enquête

La définition de la population pour laquelle les résultats du sondage doivent être mis en pratique est un aspect fondamental de la planification et de la conception de l'enquête. Si les décisions de base sur la nature et l'extension de la population à couvrir sont prises au début du processus de planification de l'enquête, le contenu et l'extension de la population doivent être spécifiés plus précisément au stade de la conception technique. Cette spécification porte sur:

- Le contenu de la population, c'est-à-dire la définition du type et des caractéristiques des unités élémentaires qui la composent;
- L'extension de la population dans l'espace, c'est-à-dire les limites de sa couverture géographique;
- Et son extension dans le temps, c'est-à-dire la période à laquelle elle se réfère.

Quelques exemples peuvent être examinés. Dans la plupart des enquêtes sur la main-d'œuvre, la population concernée est constituée par les ménages privés et l'ensemble des personnes qui y résident. Une définition de facto ou de jure peut être utilisée pour celle des membres. Dans les enquêtes sur le travail des enfants (ETE) visant à mesurer la proportion des enfants engagés dans des activités liées au travail, la population cible est constituée par tous les enfants d'un groupe d'âge donné. Dans les ETE «régulières», la couverture peut être limitée aux enfants vivant dans des ménages privés; de même, les ménages sans enfants ne font pas partie de la population cible. Dans les ETE «spéciales», la couverture peut être élargie – ou même confinée, comme dans le cas des enquêtes sur les enfants des rues – aux enfants vivant à l'extérieur de ménages privés. Dans les enquêtes sur les enfants travailleurs (EET) visant à étudier les caractéristiques et la situation des enfants engagés dans des activités en relation avec le travail, la population cible est constituée par les enfants ainsi engagés; toutefois, on complète généralement celle-ci en incluant dans l'échantillon quelques enfants non travailleurs à des fins de comparaison.

Quelques points généraux, mais importants, doivent être soulignés en ce qui concerne le choix de la population cible.

- Dans toute enquête, les règles concernant l'inclusion et l'exclusion de population doivent être définies en termes opérationnels clairement énoncés. Dans le cas contraire, confusions et erreurs en résultent au stade de la mise en œuvre.
- Les limites de la population couverte ne doivent pas être oubliées quand on tire des inférences des résultats de l'enquête et que l'on compare les résultats provenant de sources différentes.
- Mis à part les exclusions délibérées et explicites, les enquêtes souffrent aussi d'erreurs de couverture qui sont moins faciles à identifier et à mesurer. Un travail laborieux est habituellement nécessaire pour contrôler ces erreurs et en apprécier les effets sur les résultats de l'enquête. Leur ampleur dépend de la qualité de la base de sondage et de la mise en œuvre de l'échantillon.

3.3.2. La base de sondage

La population qui fait l'objet de l'enquête doit être représentée sous une *forme physique* à partir de laquelle pourront être tirés des échantillons du type requis. Une *base de sondage* est une telle représentation. Dans le cas le plus simple, la base n'est qu'une liste explicite de toutes les unités de la population, dans laquelle un échantillon des unités concernées peut être prélevé directement. Avec des plans plus complexes, la représentation de la base peut être en partie implicite, *mais elle n'en représente pas moins toutes les unité*s.

Une base peut être construite à partir d'une source unique, ou bien résulter d'une compilation obtenue par combinaison d'informations d'un certain nombre de sources. Des types et/ou des sources de bases différents peuvent être utilisés pour des parties différentes de la population. Il est également possible d'utiliser plus d'une base en combinaison afin de représenter la même population de manière plus adéquate. L'utilisation de bases multiples pose des problèmes particuliers de représentation des probabilités de sélection des unités.

Les bases de sondage par zone

Dans la pratique, la base requise est définie en relation avec la structure de l'échantillon requise et la procédure de sélection de celui-ci. Dans certains pays plus avancés, il est possible de choisir directement un échantillon de personnes sur un registre d'état-civil. Cependant, de manière plus générale, pour les enquêtes basées sur les ménages, en particulier dans les pays en développement, la base consiste généralement en un ou plusieurs stades d'unités de zone, *suivis par des listes de ménages ou d'habitations au sein d'unités finales de zone sélectionnées:*

- La base d'échantillonnage primaire (BEP) est une base d'unités d'échantillonnage primaires (UEP) et doit couvrir la population tout entière de manière exhaustive et sans recouvrements. À la suite du premier stade de la sélection, la liste des unités de tout stade ultérieur n'est requise que dans le cadre des unités plus vastes sélectionnées au stade précédent.

- Éventuellement, une hiérarchie de bases secondaires par secteur – consistant à chaque stade en l'ensemble des unités de ce stade qui existent dans chaque unité sélectionnée au stade précédent – jusqu'à l'obtention d'une base d'unités finales de secteur (UFS) peut être établie. En dessous des UFS, le processus d'échantillonnage passe des secteurs au listage et à la sélection individuelle d'habitations, de ménages ou de personnes. Dans les enquêtes auprès des ménages, la pratique courante fait intervenir un seul stade sectoriel, auquel cas les UEP sont identiques aux UFS.
- Il peut y avoir des listes explicites des unités finales d'échantillonnage (UFE) telles que les habitations ou les ménages au sein des zones d'échantillon sélectionnées.
- Les éléments d'analyse et de collecte de données de l'enquête peuvent être les UFE elles-mêmes, ou bien d'autres unités exclusivement identifiables à partir des UFE grâce à des règles d'association définies. Ainsi, des personnes (éléments) peuvent être associées à des ménages sélectionnés (UFZ) sur la base d'une définition de facto ou de jure de la couverture.

Dans la plupart des enquêtes qui ont été menées à ce jour sur la main-d'œuvre et des enquêtes indépendantes sur le travail des enfants, la base d'échantillonnage primaire a été établie à partir des zones de recensement, et le modèle de l'enquête n'a comporté qu'un seul stade sectoriel. On prépare normalement de nouvelles listes des ménages ou des habitations au sein de chaque zone sélectionnée (UEP).

La pérennité de la base diminue lorsqu'on descend dans la hiérarchie des unités pour passer des UEP aux UFE. La base d'échantillonnage primaire (et, dans une moindre mesure, les bases d'unités de niveau intermédiaire) constitue généralement un investissement majeur à long terme. Par contre, dans la plupart des enquêtes, il est nécessaire de préparer des listes actualisées des UFE peu de temps avant le dénombrement de l'enquête. Dans une enquête, un avantage majeur consiste dans la possibilité d'utiliser des listes préparées en vue de quelque autre enquête ou recensement récents, comme des listes d'EMO dans le cas d'une ETE liée à une EMO. Les listes d'unités structurelles telles que les habitations sont généralement plus durables que les listes d'unités sociales telles que les ménages; les listes préexistantes de personnes physiques sont le plus souvent de peu d'utilité, du moins dans la plupart des pays qui n'entretiennent pas correctement leurs registres d'état-civil.

Ce concept de «pérennité» différenciée des différents types d'unités est particulièrement important dans le contexte des études liées, où la seconde est basée sur un sous-échantillon de la première, comme dans le cas d'une enquête sur le travail des enfants basée sur une enquête plus vaste sur la main-d'œuvre. Des listes des différents niveaux d'unités peuvent être fournies par l'EMO et un sous-échantillon de ce niveau d'unités choisi pour l'ETE, sans autre échantillonnage en dessous de ce niveau dans l'ETE, qui se déroule indépendamment de l'EMO. Par exemple, l'échantillon de l'ETE peut être issu de (1) toutes les zones de l'échantillon d'EMO, ou (2) seulement des zones comportant des ménages avec enfants, ou (3) seulement des zones comportant des ménages avec enfants travailleurs, ou (4) de listes actualisées de tous les ménages choisis dans l'EMO , ou (5) seulement des ménages dénombrés positivement dans l'EMO, ou (6) seulement des ménages comportant des enfants, ou (7) seulement des

ménages comportant des enfants travailleurs; ou (8) de listes actualisées d'enfants listés dans les ménages de l'EMO, ou (9) de listes comportant seulement des enfants identifiés comme travailleurs et enfin (10) l'information sur les caractéristiques de ces enfants travailleurs peut être approfondie.

Il est clair que lorsque l'on suit la liste ci-dessus dans un sens descendant, la pérennité des unités impliquées dans les liens entre échantillons EMO-ETE va en diminuant, et par conséquent le décalage temporel entre les deux opérations décline comme le fait la liste ci-dessus. (Voir la section 4.4 ci-dessous pour des commentaires plus approfondis.)

Les problèmes courants posés par les bases sectorielles

Les bases par zone sont plus stables que les bases par liste. Néanmoins, elles sont également exposées aux erreurs de couverture et connexes. Celles-ci proviennent généralement du défaut de définition et d'identification des limites physiques des unités de zone, ou de la mauvaise qualité des listes d'unités finales telles que les habitations ou les ménages au sein des zones de l'échantillon. Parmi les imperfections des bases sectorielles figurent celles-ci:

- *Le défaut de couverture exhaustive de la population concernée*. Par exemple, dans un certain nombre de pays en développement où la cartographie est inadéquate, les bases disponibles sont en fait composées de listes de localités plutôt que d'unités de zone appropriées, et dans ce cas, il arrive que des populations éparpillées à l'extérieur des localités listées ne soient pas couvertes correctement. Par ailleurs, la sous-couverture augmente avec l'obsolescence de la base.
- *Les erreurs et les changements dans la délimitation sectorielle*. Ils peuvent résulter d'erreurs dans l'identification des limites sectorielles et de changements intervenus après la préparation de la base. Les limites des unités, telles qu'elles sont cartographiées ou décrites, peuvent différer des limites des unités sur lesquelles d'autres informations pertinentes sont disponibles dans la base de sondage (informations portant par exemple sur la taille et la densité de la population), ou des frontières des unités d'échantillonnage réelles.
- *L'inadéquation du type et de la taille des unités*. Les unités disponibles peuvent être trop grandes, trop petites ou de taille trop variable pour servir d'unités d'échantillonnage efficientes.
- *Le défaut d'information complémentaire*. L'information concernant la taille et autres caractéristiques des unités qui est nécessaire pour une sélection efficiente des échantillons peut être inexacte ou tout simplement indisponible.
- *Un coût élevé*. Les bases sectorielles sont généralement coûteuses à créer et à entretenir, à moins qu'elles n'existent déjà à d'autres fins d'ordre administratif. En général, l'investissement n'est justifié que quand la base doit être utilisée de manière répétée pour des enquêtes et degrés d'enquête multiples.

Les problèmes des bases par liste

Des problèmes peuvent surgir en l'absence de correspondance point par point entre les *listings* (qui sont les unités réellement soumises au processus de sélection) et les *unités*

élémentaires (desquelles l'objectif réel étant d'en obtenir un échantillon avec des probabilités spécifiées). Ce défaut de correspondance peut revêtir plusieurs formes.

Les entrées en blanc:sur une liste sont celles qui ne représentent pas une unité réelle. La présence de blancs sur une liste n'affecte pas les probabilités de sélection des unités, mais le nombre des unités sélectionnées devient une variable aléatoire. Si ce nombre est maintenu fixe, les probabilités de sélection sont sujettes à variation aléatoire et deviendront inconnues si le nombre d'unités réelles représentées sur la liste n'est pas connu.

Des blancs sur une liste *ne représentent pas* des non-réponses. Une erreur courante consiste à traiter les blancs comme des non-réponses et à les remplacer par d'autres unités ou à ajuster la pondération pour les compenser.

Les duplications: apparaissent quand une même unité est représentée par plus d'une entrée. Parfois, le problème est dû à la nature de la base - par exemple, dans la sélection de ménages sur un registre électoral (listant toutes les personnes admises à voter dans chaque ménage), dans la sélection d'un parent sur une liste d'enfants scolarisés, ou dans la sélection de clients ou de bénéficiaires de services sur des registres de visites dans une installation de services. Généralement, la sélection d'unités sujettes à des événements multiples sur la base d'un listage d'événements distincts confère à chaque unité une probabilité de sélection proportionnelle au nombre d'événements auquel elle est associée. Il convient de prendre des mesures pour compenser ou éviter ces variations des probabilités de sélection: *la simple élimination des duplications qui apparaissent dans l'échantillon ne résout pas le problème*.

Beaucoup plus difficile à résoudre est le problème des duplications non systématiques sur une liste, qui résulte d'un défaut d'identification du fait que des entrées différentes représentent en réalité la même unité . Cela peut se produire, par exemple, si la même unité est enregistrée plusieurs fois sur la même liste avec de légères différences de nom, d'adresse ou de description. Dans des cas semblables, un travail méticuleux d'élimination de toutes les duplications peut être la seule solution.

La mise en grappes des éléments: signifie que la même entrée peut représenter plus d'une unité. En soi, cela ne distord pas les probabilités de sélection, puisque chaque unité reçoit la probabilité de sélection de l'entrée qui la représente. *Souvent, la sélection aléatoire d'une seule unité dans la mise en grappes n'est pas nécessaire, mais elle peut être inévitable, voire souhaitable, dans certaines situations.* Un exemple courant en est la sélection d'un seul adulte de chaque ménage échantillonné en vue de son inclusion dans l'enquête. *Si l'on procède ainsi, les résultats doivent être pondérés pour refléter les changements de probabilité de sélection.* Avec un tel modèle, une unité (par exemple des personnes) reçoit une probabilité proportionnelle au nombre d'unités représentées dans son entrée (ménage).

La sous-couverture: signifie que certaines unités ne sont pas représentées dans la base. C'est là le problème le plus sérieux et le plus difficile, qui fausse les résultats de nombreuses enquêtes. Cela provient du fait que la non-représentation dans la base n'est généralement pas aléatoire, mais *sélective* en fonction des caractéristiques des unités. Il n'existe pas de solution simple et peu coûteuse au problème de la sous-couverture. Le

seul conseil que l'on puisse donner est de consacrer davantage d'efforts à la préparation ou à la compilation des listes.

Le défaut de localisation des unités: consiste à ne pas identifier correctement la/les unité(s) que représente une entrée sélectionnée. Le problème est courant en l'absence d'une description claire et complète dans la base permettant d'identifier les unités sur le terrain (noms des enfants, par exemple). Il peut aussi résulter de l'insuffisance des efforts des agents d'exécution. *Le défaut de localisation d'unités sélectionnées qui existent réellement constitue une non-réponse.* Ce problème est souvent confondu avec celui des blancs, qui concerne des unités sans existence réelle (voir ci-dessus). Les unités non localisées sont souvent signalées sans discernement comme non existantes.

3.4. Les écarts par rapport au sondage aléatoire simple: stratification, échantillonnage par grappes et probabilités inégales

3.4.1. Les écarts par rapport au sondage aléatoire simple

La conception la plus simple est celle dans laquelle tout ensemble possible de s=1 à n unités dans une population de N unités reçoit la même chance de sélection. On désigne cet ensemble sous l'appellation d'échantillon aléatoire simple (EAS). Il existe

$$\frac{N!}{(N-n)!n!}$$

échantillons de ce type, et chacun reçoit une probabilité de sélection inverse au nombre ci-dessus. En fait, comme on l'a noté, dans un EAS, tout ensemble de s, $1 \leq s \leq n$, unités reçoit la même chance d'appartenir à l'échantillon que tout autre ensemble de la même taille. Des unités différentes (ce qui correspond à s=1) reçoivent toutes la même chance, soit n/N. D'autres modèles s'écartent de l'échantillonnage aléatoire simple en:

1. supprimant certains des échantillons possibles indiqués ci-dessus, autrement dit en ne permettant pas à certaines combinaisons d'unités d'apparaître dans le même échantillon, et/ou

2. donnant à des unités différentes (donc à des échantillons différents) des probabilités de sélection différentes.

La conception de l'échantillon peut s'écarter de l'échantillonnage aléatoire simple d'un certain nombre de manières, dont les plus courantes et les plus importantes sont les trois suivantes.

La stratification

La stratification consiste à *fractionner la population* avant la sélection de l'échantillon. Dans chaque partie obtenue, ou strate, un échantillon est sélectionné séparément (indépendamment). Dans chaque strate, la conception peut comporter d'autres démarches complexes, comme l'échantillonnage par grappes ou l'échantillonnage à

plusieurs degrés, et elle peut différer d'une strate à l'autre. Les principaux objectifs de la stratification sont de gagner en flexibilité dans la conception de l'échantillon et dans son allocation à différentes parties de la population et d'accroître l'efficience statistique du plan d'échantillonnage. (Parmi tous les échantillons possibles dans le cadre de l'EAS, seuls ceux qui contiennent un nombre fixe ou attendu d'unités de chaque strate peuvent résulter de ce plan d'échantillonnage.)

L'échantillonnage par grappes ou l'échantillonnage à degrés multiples

De même que la stratification évoque le fractionnement de la population avant la sélection de l'échantillon, l'échantillonnage par grappes évoque le groupement d'unités.

Il est souvent économique et commode de grouper les éléments de population en unités plus importantes («grappes») et d'appliquer les procédures de sélection à ces groupes plutôt que de manière directe aux unités élémentaires individuelles. En pratique, cet échantillonnage par grappes est souvent en fait la seule option disponible, car les éléments individuels sont trop nombreux et trop éparpillés pour être échantillonnés directement. (Parmi tous les échantillons possibles dans le cadre de l'EAS, seuls ceux qui comportent les unités finales limitées aux unités sélectionnées lors des stades précédents peuvent résulter de ce plan d'échantillonnage.)

La procédure de sélection peut être plus élaborée que la simple sélection d'un ensemble de grappes. Par exemple, quelques grandes unités peuvent être sélectionnées d'abord; ensuite, chaque unité sélectionnée pourra être divisée en unités plus petites, et un échantillon de celles-ci sera choisi; et finalement, dans chacune des unités plus petites qui auront été choisies, un échantillon d'éléments individuels pourra être sélectionné. Nous obtiendrons ainsi un *plan à plusieurs degrés*. L'objectif de ce type de plan est de confiner les éléments qui apparaissent dans l'échantillon aux unités plus importantes sélectionnées au(x) stade(s) précédent(s). On procède habituellement ainsi pour réduire le coût des enquêtes et améliorer le contrôle de l'opération de collecte de données de l'enquête.

Les probabilités de sélection inégales

Il existe parfois des raisons de sélectionner des classes d'éléments avec des probabilités supérieures (ou inférieures) à d'autres. Par exemple, certains domaines, c'est-à-dire des parties de la population comme les zones urbaines ou les petites régions d'un pays, peuvent être suréchantillonnés afin d'améliorer la précision de leurs résultats. Cela peut se révéler particulièrement nécessaire dans les enquêtes sur le travail des enfants quand la population concernée est inégalement distribuée dans le pays. Habituellement, des probabilités de sélection inégales exigent une pondération des données de l'échantillon au stade de l'estimation. Cette pondération peut aussi être introduite pour d'autres raisons, par exemple pour compenser les non-réponses, ou pour calibrer les (ou certaines) caractéristiques de l'échantillon pour répondre à des critères extérieurs (voir chapitre 7).

Les incidences sur la variance

L'incidence de ces écarts sur la variance par rapport à l'échantillonnage aléatoire simple peut être considérée comme suit. Avec un plan EAS d'une taille donnée n, toutes les

combinaisons possibles de tout nombre donné s, $1 \leq s \leq n$, éléments ont des chances égales d'apparaître dans l'échantillon. L'échantillonage par grappes ou la stratification ont pour effet de supprimer certaines de ces combinaisons possibles (et par conséquent à supprimer le sous-ensemble d'échantillons qui en contient une quelconque). Les probabilités inégales de sélection des unités rendent certains résultats d'enquête plus vraisemblables que d'autres.

L'échantillonage par grappes tend à supprimer un nombre relativement plus élevé de bons échantillons (autrement dit, des échantillons donnant des statistiques proches des vrais paramètres de population qui sont soumis à estimation) et à conserver davantage de «mauvais» échantillons donnant des résultats plus éloignés des véritables valeurs de population. En conséquence, la variabilité entre les échantillons conservés tend à être plus importante, ce qui signifie que l'échantillonnage par grappes amplifie l'erreur d'échantillonnage.

C'est l'opposé qui se produit avec la stratification, qui tend à supprimer une proportion relativement plus importante de «mauvais» échantillons et à conserver davantage de «bons» échantillons. En conséquence, la variabilité entre ceux qui ont été conservés tend à être plus faible, ce qui signifie que la stratification réduit l'ampleur de l'erreur d'échantillonnage.

L'incidence des probabilités de sélection inégales est plus complexe. Quand ces variations sont essentiellement aléatoires ou n'ont qu'une faible corrélation avec les caractéristiques des unités, elles introduisent habituellement une variabilité supplémentaire au sein des résultats du sondage. Toutefois, des techniques telles que l' «allocation optimale» et le «calibrage», qui introduisent des taux de sélection ou de pondérations inégaux, peuvent parfois réduire effectivement la variance.

3.4.2. La stratification: quelques aspects pratiques

Le but de la stratification

La stratification consiste à diviser les unités de la population en groupes, puis à choisir indépendamment un échantillon au sein de chaque groupe, ce qui assure un contrôle séparé du plan et de la sélection de l'échantillon au sein de chaque strate. Cela veut dire que les segments de la population (strates) peuvent être échantillonnés différemment, grâce à l'utilisation de taux et de plans d'échantillonnage différents. Bien que ce ne soit pas un aspect essentiel du concept de stratification, la séparation peut aussi être maintenue au stade de la mise en œuvre, de l'estimation et de l'analyse. Il est courant, par exemple, de centraliser les résultats provenant de différentes strates pour produire des estimations concernant la population tout entière, ou bien des parties essentielles ou des «domaines» de la population, dont chacun est composé d'un certain nombre de strates. Les avantages de la stratification proviennent du contrôle qu'elle permet d'exercer sur le plan et la sélection des échantillons au sein de chaque strate:

- Premièrement, dans la mesure où les strates représentent des groupements d'unités relativement homogènes, l'échantillon qui en résulte gagne en efficience dès lors que les unités appartenant à chaque groupement sont représentées de manière adéquate et contrôlée.

- Quand il faut disposer, avec une précision spécifiée, de données séparées pour des subdivisions de la population, il est souhaitable de traiter chacune des subdivisions comme une «population» en soi et de sélectionner indépendamment dans chacune d'elles un échantillon de la taille et du plan requis: la stratification le permet.
- Les conditions et les problèmes de l'échantillonnage – pour ce qui est de la taille et du plan de l'échantillon, la disponibilité d'une base pour la sélection des échantillons, les conditions de voyage, etc. – peuvent différer sensiblement entre des parties différentes de la population. La stratification autorise une certaine flexibilité dans le choix de plans séparés au sein de chaque partie.
- Un échantillon clairement contrôlé et distribué de manière proportionnelle (ou selon d'autres critères spécifiés) entre les différentes parties de la population présente, en termes de relations publiques, l'avantage d'apparaître comme plus «représentatif», donc plus acceptable pour les utilisateurs. Dans tous les cas, le contrôle exercé grâce à la stratification réduit les risques de mauvaise distribution des échantillons.
- La stratification peut aussi être introduite pour des raisons de commodité administrative; ainsi, par exemple, la sélection et la mise en œuvre des échantillons peuvent être confiées à des bureaux extérieurs différents, dont chacun s'occupera de sa propre «strate».

La stratification dans la pratique

En pratique, la stratification d'une liste ou d'une base avant la sélection de l'échantillon justifie en général des soins et des efforts considérables, et cela pour les raisons suivantes:

La stratification réduit souvent la variance d'échantillonnage pour un coût additionnel limité. En outre, les coûts tendent à être plus faibles et les avantages plus importants en cas de stratification d'unités d'un stade plus élevé dans un plan à plusieurs degrés, par comparaison avec les avantages de la stratification d'unités d'un degré inférieur ou d'un échantillon d'éléments (voir ci-dessous). Il est souvent souhaitable de poursuivre la stratification jusqu'à sa limite, où seules une ou deux UEP sont sélectionnées par strate. En fait, des techniques spéciales connues sous l'appellation de «sélection contrôlée» peuvent être employées pour créer un nombre de strates encore plus grand que le nombre d'unités à sélectionner, en liant les sélections de différentes strates afin de réaliser la répartition de l'échantillon requise.

- Dans la mesure où les échantillons sont sélectionnés de manière indépendante, et où leur taille est suffisante, les résultats de chaque strate peuvent être analysés et présentés séparément. Plus couramment, les résultats de plusieurs strates sont agrégés pour produire des estimations relatives aux grands domaines de la population. On améliore l'efficience en définissant les strates dans les limites des domaines de rapport (sans les dépasser).
- Une utilité très importante de la stratification est de favoriser la flexibilité dans le choix de la répartition de l'échantillon, de sa conception et des procédures utilisées dans différentes parties de la population.

- Les strates peuvent offrir des séparations naturelles pour organiser, contrôler et échelonner le travail d'enquête. Généralement, la stratification ne complique aucunement les opérations sur le terrain au stade de la collecte de données. Au contraire, toute complexité additionnelle est confinée à l'opération de sélection de l'échantillon, qui est habituellement plus centralisée et par conséquent plus facile à contrôler.
- L'échantillonnage systématique à partir de listes organisées est un moyen peu onéreux et efficient de réaliser les effets de la stratification. Cette procédure tend à être beaucoup plus simple à mettre en œuvre que la sélection avec utilisation de nombres aléatoires.

La stratification dans l'échantillonnage à plusieurs degrés

Les arguments en faveur d'une stratification attentive et minutieuse deviennent beaucoup plus convaincants quand nous examinons les plans à plusieurs degrés.

- Le point essentiel est que le gain de précision au moyen de la stratification est généralement beaucoup plus important dans l'échantillonnage à plusieurs degrés que dans l'échantillonnage par éléments.
- On dispose généralement d'informations beaucoup plus nombreuses pour la stratification des unités plus importantes, telles que les zones de recensement ou les localités faisant office d'UEP et autres unités d'un degré supérieur dans un plan à plusieurs degrés.
- Il est plus facile de stratifier les unités plus importantes et d'un stade supérieur, qui tendent à être beaucoup moins nombreuses que le nombre d'éléments de la population.
- Dans la mesure où le nombre des unités d'un degré supérieur est faible, il devient important de veiller à ce que la répartition de l'échantillon soit contrôlée. Pour cela, on procédera à un échantillonnage séparé à l'intérieur des strates.
- Dans l'échantillonnage à plusieurs degrés, il est plus souvent nécessaire, et aussi plus réalisable, de faire varier la procédure d'échantillonnage selon les parties de la population.

Les critères de stratification

Nous présentons ci-dessous un certain nombre de suggestions utiles pour le choix des critères utilisés pour la stratification.

- La stratification s'effectuant avant la sélection des échantillons, on peut faire des choix subjectifs pour déterminer les critères de définition, le nombre et les limites des strates. L'uniformité (l'utilisation de la même procédure dans toutes les strates) et l'objectivité (le recours à des critères et procédures prédéterminés ne faisant pas intervenir le jugement) ne sont pas exigés à ce stade, ce qui contraste avec la nécessité de suivre des procédures objectives au stade de la sélection proprement dite pour établir un échantillon aléatoire.
- Il est généralement plus efficace d'utiliser une multiplicité de variables de stratification, comportant chacune quelques catégories, que d'utiliser un grand nombre de catégories détaillées d'une seule variable.

- La stratification par unité de taille est utile quand cette taille varie largement. Un des objectifs de cette stratification est de contrôler la taille des échantillons, bien que des procédures spéciales telles que l'échantillonnage avec probabilité proportionnelle à la taille (PPT) puissent aussi permettre ce contrôle dans une certaine mesure. Un autre objectif de la stratification par taille est de contrôler la répartition des caractéristiques qui sont liées à la taille de l'unité. Cela est utile même quand l'échantillonnage avec PPT a été utilisé pour contrôler les prélèvements d'échantillons au sein des grappes. (Voir dans la section 3.6 les explications relatives à la méthode d'échantillonnage avec PPT.)
- Dans de nombreuses situations, la classification géographique, administrative et «urbain-rural» constitue la forme de stratification la plus efficace. Ces formes de stratifications sont simples et demandent peu d'informations complémentaires. Elle tend aussi à s'adapter à des enquêtes portant sur des thèmes divers.
- Ce qui précède s'applique plus clairement à la stratification d'unités d'un degré supérieur dans les plans à plusieurs degrés. Les caractéristiques individuelles (telles que la taille des ménages, ou le sexe et l'âge des personnes) peuvent bien entendu être des critères de stratification efficace supplémentaires dans les cas d'échantillonnage d'unités individuelles.

Le type de stratification le plus communément utilisé pour la sélection des UEP et autres zones dans les enquêtes basées sur les ménages portant sur la population générale est géographique: stratification selon le type de lieu (urbain-rural, ou plusieurs catégories classées par degré d'urbanisation ou taille de la localité); localisation (province, région ou autre division administrative) et zone climatique ou écologique. Des systèmes de stratification plus complexes sont habituellement nécessaires pour l'échantillonnage d'unités à caractéristiques spéciales, en particulier quand les unités concernées sont inégalement distribuées au sein de la population.

À l'intérieur des zones d'enquête, les ménages peuvent être stratifiés en fonction de leur taille, de leur situation socio-économique, de l'emploi du chef du ménage, etc., dans la mesure où ce type d'information est disponible. Cependant dans la pratique, cette stratification au niveau du ménage (ou au niveau personnel) peut imposer des exigences considérables de données détaillées, et par conséquent être coûteuse. De plus, son efficacité en matière de réduction de la variance peut en fait être très limitée. Dans les échantillons à degrés multiples les plus typiques, les gains les plus importants proviennent plutôt de la stratification aux stades supérieurs.

3.4.3. L'échantillonnage par grappes

Les raisons de l'échantillonnage à degrés multiples

Dans certaines situations et à certaines fins, la sélection directe d'éléments en un seul stade peut être plus simple et plus efficiente que la sélection d'un échantillon par degrés multiples. Toutefois, dans la plupart des circonstances, notamment pour les enquêtes à grande échelle menées auprès des ménages dans les pays en développement, la sélection directe des éléments (ménages ou personnes) n'est pas une option réalisable. L'échantillonnage à degrés multiples est introduit pour un certain nombre de raisons.

- En concentrant en grappes les unités à dénombrer, il réduit les frais de voyage et autres coûts de la collecte de données.
- Pour la même raison, il peut améliorer la couverture, la supervision, le contrôle, le suivi et d'autres aspects qui déterminent la qualité des données collectées.
- La commodité administrative dans la mise en œuvre de l'enquête peut constituer une autre raison importante.
- Le fait de choisir l'échantillon en plusieurs degrés successifs réduit le travail et les coûts impliqués par la préparation et l'entretien de la base de sondage. Premièrement, avec l'échantillonnage à degrés multiples, il suffit de disposer d'une base couvrant l'ensemble de la population pour la sélection des UEP dans un premier stade; à tous les degrés inférieurs, une base est nécessaire seulement au sein des unités sélectionnées au degré précédent. Par contre, pour l'échantillonnage direct d'éléments tels que les ménages ou les personnes, une liste complète couvrant tous les éléments de la population sera requise. Deuxièmement, lorsque l'unité concernée est d'une taille importante, les bases tendent à être plus durables et par conséquent utilisables sur de plus longues périodes; les listes de petites unités comme les ménages et particulièrement les personnes tendent à devenir rapidement obsolètes.
- Le travail qu'implique la sélection de l'échantillon peut également être réduit par l'échantillonnage à degrés multiples. Ainsi, il est plus facile de traiter quelques centaines ou quelques milliers d'unités de zone dans un pays que de choisir un échantillon sur des listes de millions de ménages. Il est plus facile de classifier et de stratifier des unités plus importantes, et l'on dispose habituellement d'informations beaucoup plus nombreuses à cet effet.

Le coût de l'échantillonnage par grappes

Les avantages ci-dessus doivent être comparés aux coûts divers de l'échantillonnage à degrés multiples.

- Le coût majeur de l'échantillonnage par grappes ou de l'échantillonnage à degrés multiples correspond à l'augmentation de l'erreur d'échantillonnage par rapport à celle d'un simple échantillon aléatoire de la même taille (c'est-à-dire avec le même nombre d'éléments dénombrés). L'augmentation de la variance varie selon l'homogénéité relative des éléments au sein des unités de stades supérieurs, et du mode de sélection et des unités sélectionnées à chaque stade. Si les éléments regroupés au sein d'une unité de stade supérieur sont assez semblables entre eux, chacune des unités donne, en un sens, moins d'informations nouvelles que l'on n'en obtiendrait si tous les éléments étaient choisis de manière aléatoire dans la population tout entière. Cela tend à limiter l'efficience de l'échantillon. La perte d'efficience sera plus importante si le nombre d'éléments choisis par grappe est augmenté, ou si les éléments sont plus étroitement regroupés en unités compactes, ou si les éléments d'une même grappe tendent à être plus homogènes quant à la variable d'intérêt.
- Il peut aussi y avoir une certaine perte de flexibilité dans la conception de l'échantillon et dans le ciblage de l'échantillon des populations aux caractéristiques particulières. En effet, des éléments de types différents sont

généralement mêlés au sein des unités de stade supérieur, si bien qu'il est impossible de contrôler séparément la sélection d'unités de n'importe quel type donné.

- La complexité du plan augmente aussi la complexité de certains types d'analyses des données d'enquête. Cela s'applique en particulier à l'estimation des erreurs d'échantillonnage, qui doit prendre en compte la structure de l'échantillon.

Dans les pays qui sont statistiquement plus développés, la balance penche de plus en plus en faveur de l'échantillonnage direct des éléments, ou du moins d'une réduction de l'échantillonnage par grappes des échantillons dans les enquêtes auprès des ménages. Les facteurs qui y contribuent sont:

- le coût accru du temps nécessaire pour obtenir réellement et analyser les données d'entretien, comparé aux frais de voyage;
- les nouveaux modes de collecte de données, comme l'entretien téléphonique;
- la nécessité de rendre les échantillons plus efficients, afin de pouvoir réduire leur taille; et
- la disponibilité de meilleures bases permettant une sélection directe des éléments.

Le choix du type d'unité qui servira de zone d'enquête

Avec l'échantillonnage à degrés multiples, le choix du type d'unité sectorielle qui sera utilisé dans l'enquête, ainsi que le nombre de ces unités qui seront sélectionnées pour l'échantillon, sont importants.

Le choix du type d'unité qui servira de zone d'enquête a des implications majeures, car le type d'unité choisi pour servir d'UEP et autres unités de stade élevé peut affecter considérablement la qualité, le coût et le fonctionnement de l'enquête.

Tout d'abord, quelques conseils généraux:

- Il n'est ni nécessaire ni toujours efficient, d'insister sur l'utilisation d'unités du même type ou de la même taille que les UEP dans tous les domaines de population à échantillonner.
- Il est très courant que des types d'unités différents portent la même étiquette administrative. Il est important de ne pas confondre les étiquettes administratives formelles avec le type réel d'unité concerné.
- Le type et la taille des unités appropriés dépendent des circonstances et des objectifs de l'enquête.
- Néanmoins, le choix est soumis à de sérieuses contraintes imposées par ce qui est disponible dans la base de sondage.
- Le choix approprié dépendant des circonstances, aucune pratique standard ou unique ne peut être recommandée.

Manifestement le type d'unités sélectionnées servant d'UEP peut grandement influencer sur la qualité et le coût et l'enquête. En ce qui concerne les zones d'enquêtes, les unités pertinentes doivent être bien définies, avec des limites évidentes. De bonnes cartes et

descriptions, assorties d'informations actualisées sur la taille et les caractéristiques, sont nécessaires à l'identification et à la démarcation de ces zones. Ces dernières doivent couvrir la population entière de manière exhaustive et sans chevauchements. La stabilité à travers le temps est une autre condition importante, surtout si les unités sont utilisées sur une longue période. Les UEP doivent être de taille «appropriée», en rapport avec l'organisation et la structure des coûts de l'opération de collecte de données de l'enquête. Si les unités sont trop grandes, il ne sera pas possible d'inclure un nombre suffisant d'entre elles pour obtenir une bonne couverture de l'échantillon. De plus, les coûts du listage, du sous échantillonnage et de la collecte de données pour des grandes unités peuvent devenir excessifs. D'autre part, si les unités sont trop petites et compactes, il peut être difficile d'assurer une étendue suffisante des unités afin d'obtenir un échantillon performant. Les petites unités de zones ont également tendance à avoir des limites peu définies et une moindre stabilité dans le temps. Par ailleurs, plus les unités sont petites, plus nous en avons besoin pour réaliser l'échantillon final.

Ce qui constitue une «taille appropriée» pour des unités servant d'UEP dépend des circonstances et des objectifs de l'enquête. Des considérations pratiques diverses doivent simultanément retenir l'attention; parmi celles-ci figurent le coût, le contrôle de qualité, l'administration, la disponibilité de la base pour la sélection de l'échantillon, et l'efficience du plan qui en résulte. Une compréhension approfondie de ces considérations variées présuppose une solide connaissance de la théorie de l'échantillonnage et une longue pratique.

En fait, plusieurs plans peuvent être identifiés à partir de la diversité de la conduite d'enquêtes nationales dans les pays développés comme dans les pays en développement.

- Dans de nombreux pays en développement, les grands réseaux routiers sont suffisamment développés pour faciliter les voyages entre les zones, mais les déplacements locaux peuvent être plus difficiles et demander davantage de temps, même si les distances physiques à parcourir sont limitées. Cela exclut l'utilisation, dans de telles situations, d'un mode souvent utilisé dans les enquêtes des pays développés, qui fait intervenir des enquêteurs hautement mobiles, dont chacun couvre une UEP très vaste et étendue. D'un autre côté, en l'absence de bonnes cartes et autres matériels permettant de définir de petites unités de zone appropriées, l'utilisation de nombreuses petites UEP est également exclue. Le nombre des UEP doit également être limité, afin de contrôler les coûts des voyages et de la supervision. En fait, une condition courante dans le choix du plan est d'assurer que chaque UEP produise un échantillon assez vaste pour occuper les enquêteurs assez longtemps (par exemple quelques jours) dans chaque zone. Cette condition devient encore plus contraignante quand les enquêteurs sont déployés en équipes.
  Ainsi donc, de nombreuses enquêtes menées dans des pays en développement utilisent des zones de dénombrement (ZD) comme les UEP, une ZD qui se compose généralement de 100-300 ménages.
- Par contre, dans les zones urbaines de nombreux pays en développement (notamment en Amérique latine), ainsi que dans les enquêtes auprès des ménages dans les pays développés en général, de nombreux échantillons sont

basés sur des UEP nombreuses et très petites. Chaque UEP peut être une petite grappe de ménages (de 5 à 10 ménages), dont certains ou la totalité sont inclus dans l'échantillon. Un tel système peut être approprié dans les zones urbaines à forte densité de population, où l'on dispose de listes d'unités de logement et/ou de cartes très détaillées de petits segments de zone et où, du fait du développement des services de transport et des courtes distances à parcourir, les voyages entre unités au sein de la même localité ne présentent pas de problèmes particuliers.

- Ces activités, en tant qu'enquêtes intensives sur des zones ou des types de travail des enfants particuliers, peuvent rencontrer des contraintes pratiques qui imposeront de limiter l'enquête à un très petit nombre de sites. Si des études basées sur des échantillons limités peuvent fournir des informations utiles à des fins spécifiques, ce type de disposition est généralement inadéquat pour les enquêtes visant à produire des estimations statistiquement précises.

Le nombre de zones échantillonnées (ou prélèvement d'échantillons par zone)

Un examen des types d'échantillons utilisés jusqu'ici dans les enquêtes sur le travail des enfants montre l'existence d'une variation étonnamment large dans la taille des grappes utilisée (le prélèvement d'échantillons par zone). L'étendue de la variation dépasse les 10-50 ménages par zone. Si celle-ci reflète dans une certaine mesure des différences dans la situation nationale et dans le type d'unités considéré, il est probable qu'une large part de la variation ne provient pas des statistiques réelles ou sur des différences de coût.

Sur la base d'informations sur les erreurs d'échantillonnage et sur les effets du plan de sondage, et sur la répartition des coûts sur le terrain entre les zones et à l'intérieur de celles-ci, nous pouvons porter l'appréciation générale suivante sur la qualité des plans utilisés jusqu'ici dans les enquêtes nationales sur le travail des enfants:

> *Très souvent, les plans de sondage utilisés pour les enquêtes sur le travail des enfants ont été inefficients: il est nécessaire (et possible) d'améliorer leur efficience.* Cela n'est réalisable que sur la base d'une évaluation plus attentive de ce qu'implique le choix du plan en matière de coûts des enquêtes, de variance et autres aspects de la qualité des données. Cela est essentiel, même si l'évaluation de ces aspects ne peut être qu'approximative et partielle.

3.4.4. L'inégalité dans les probabilités de sélection et dans la pondération

Il peut y avoir des raisons diverses d'introduire des probabilités de sélection inégales (ou leur inverse, des pondérations d'échantillon) pour les unités finales d'un échantillon. Nous pouvons les diviser en trois types principaux.

1. Des domaines de rapport ou des sous-groupes de population différents peuvent être échantillonnés selon des taux différents pour répondre à des conditions redditionnelles spécifiques. On en trouve des exemples dans le suréchantillonnage de petits domaines visant à assurer une taille d'échantillon adéquate à l'établissement de rapports séparés par domaine, ou encore, de manière semblable, dans le suréchantillonnage de populations restreintes ou particulières

ou d'événements rares. On peut y voir des écarts «nécessaires» par rapport aux taux d'échantillonnage uniformes, écarts dictés par les objectifs fondamentaux de l'enquête.

2. Il existe aussi des divergences «inutiles» en ce sens qu'elles ne sont pas dictées par les objectifs de l'enquête mais qu'elles résultent des procédures d'échantillonnage particulières adoptées, des déficiences de la base de sondage ou d'autres circonstances de l'enquête. Ainsi, il peut exister dans des listes de multiples occurrences d'unités qui induisent ensuite des probabilités de sélection différentes, mais qui peuvent seulement être découvertes après sélection de l'échantillon. Il est souhaitable d'éviter ce type de variations, même si ceci n'est pas toujours possible dans la pratique.

3. Troisièmement, certains types de variation des probabilités d'échantillonnage peuvent être introduits pour renforcer l'efficience du plan de sondage (réduire la variance et/ou les coûts). On en trouve un exemple dans la «répartition optimale», qui consiste à suréchantillonner des strates contenant des unités qui sont plus diverses et/ou moins coûteuses à dénombrer. De même, des pondérations de «calibrage», qui rendent l'échantillon conforme aux totaux et aux répartitions d'un contrôle externe plus fiable, peuvent être introduites pour réduire la variance et les biais d'échantillonnage.

Les variations des probabilités de sélection des unités des types 1 et 2 sont externes et, en ce sens, essentiellement arbitraires, sans connexion avec les niveaux et les variances de la population. Ces variations augmentent généralement l'erreur d'échantillonnage de l'échantillon total. Leur effet est assez uniforme entre les différents types de statistiques, de variables et de sous-groupes de la population. L'accroissement des erreurs d'échantillonnage est déterminé par le coefficient de variation des probabilités d'échantillonnage ou les pondérations associées des unités de la population.

Par contre, les variations du type 3 *peuvent* occasionner une réduction de la variance ou du biais, mais cela ne se produit pas nécessairement dans tous les cas, ni de manière significative.

3.5. Le choix de la taille d'échantillon

3.5.1. La question de la taille d'échantillon

Dans toute enquête, le choix de la taille d'échantillon est manifestement la question la plus fondamentale et la plus importante, et c'est pourtant une question difficile, à laquelle on ne peut apporter une réponse purement quantitative. Le choix de la taille d'échantillon doit établir un équilibre entre

- Ce qui est *exigé* du point de vue de la précision de l'échantillonnage,
- Ce qui est *réalisable* du point de vue de l'application pratique (ex.: budget, personnel de terrain et administratif, ressources techniques, contrôle de qualité, contraintes temporelles, gérabilité, pérennité).

Le choix de la taille d'échantillon est déterminé par un équilibre entre diverses considérations statistiques, pratiques et économiques. Cet équilibre pose un problème complexe. Cette section examine, en termes généraux, les considérations qui entrent en jeu, afin de fournir des principes directeurs utiles pour le choix de la taille d'échantillon dans des situations spécifiques.

L'«échelle» d'une opération d'enquête

Il convient d'abord de réaliser que l'échelle d'une enquête n'est pas seulement une affaire de taille d'échantillon. Par «échelle», il faut entendre la mesure de la charge totale qu'engendre une enquête: celle-ci va au-delà de la taille d'échantillon et l'inclut. Elle augmente aussi en fonction de la complexité et du volume de l'information collectée pour chaque cas, de la sensibilité et de la lourdeur de l'information collectée pour les répondants (et pour l'enquêteur), de la complexité et des exigences des opérations d'enquête, etc.

Déterminer l'échelle d'une enquête est une démarche fondamentale, qui dépend de presque tout, sur lequel elle influe à son tour: mode de collecte des données, planification et organisation des opérations, pertinence, opportunité et exactitude des données obtenues et par-dessus tout, coût de l'enquête. Les observations suivantes concernant la taille d'échantillon s'applique en fait à ce large concept d'échelle des opérations d'enquête.

La nécessité d'un compromis

Il est utile de commencer par écarter des approches extrêmes ou unilatérales de la question de la taille d'échantillon, même si chacune d'entre elles peut présenter quelque intérêt dans des circonstances précises.

- Une pratique courante consiste à identifier d'abord une seule estimation cruciale, ou tout au plus quelques-unes, et à choisir un niveau plus ou moins arbitraire de précision souhaitée pour calculer la taille d'échantillon requise. Ces calculs en eux-mêmes sont généralement simples, mais très rarement réalistes, car la plupart des enquêtes ont des objectifs complexes et multiples qui ne peuvent pas être réduits à l'estimation d'un unique chiffre «crucial».
- Les objectifs de précision ne peuvent pas non plus être considérés comme entièrement prédéterminés: ils sont souvent modifiés, plus ou moins radicalement, par ce qui est réalisable compte tenu des contraintes économiques, temporelles, administratives et techniques. Les conditions à remplir en matière de précision peuvent rarement être exprimées avec exactitude ou avec une objectivité suffisante. Même s'ils sont admis, les calculs simples conduisent souvent à des exigences en matière de taille d'échantillon dénuées de réalisme, ce qui occasionne une révision arbitraire des conditions de précision initiales (tout aussi arbitraires).
- Selon une autre position, le choix de la taille d'échantillon est avant tout une question non technique, déterminée presque exclusivement par des considérations de coût et autres considérations pratiques, l'expérience antérieure (sinon les préjugés), le désir de parité avec d'autres enquêtes et

d'autres pays et, de manière générale, la volonté de s'assurer la taille d'échantillon maximale accessible dans les circonstances données.

- On soutient couramment que les erreurs extérieures à l'échantillonnage prédominent généralement sur les erreurs d'échantillonnage dans les enquêtes à grande échelle. Cette assertion peut paraître assez pratique, mais elle n'est ni significative ni utile quand on l'exprime en ces termes absolus. En effet, elle reviendrait simplement à l'affirmation assez extraordinaire selon laquelle les tailles d'échantillon utilisées dans la pratique seraient trop importantes *en général.* Premièrement, il y a toujours, dans la pratique, un compromis entre l'ampleur des erreurs d'échantillonnage et celle des autres erreurs. Deuxièmement, les erreurs d'échantillonnage deviennent toujours la composante prédominante lorsque l'on produit des estimations applicables à des domaines de plus en plus petits.

L'approche appropriée est une combinaison équilibrée de ces perspectives, qui évite les visions simplistes du problème complexe posé par la taille d'échantillon.

Les conditions minimales

Il se produit effectivement des situations dans lesquelles, au moins au point de départ, la taille d'échantillon est déterminée essentiellement sur la base des conditions de précision de quelques statistiques «cruciales» ou fondamentales. Toutefois, les conditions de précision sont habituellement beaucoup plus complexes, même si quelques variables cruciales seulement sont prises en compte. La classification en tableaux et l'analyse des résultats d'enquête impliquent de nombreux types d'estimations pour des sous-populations diverses. *Dans la plupart des situations, on prend d'abord en considération le degré de détail avec lequel les données de l'échantillon peuvent être classifiées de manière significative.* Toutefois, au-delà d'un certain minimum au-dessous duquel l'enquête peut perdre son utilité, ce degré n'est pas entièrement prédéterminé. En fait, il est lui-même conditionné par ce qui est possible avec différents choix de taille d'échantillon.

En dépit de la complexité des conditions fondamentales, il est possible d'identifier les situations dans lesquelles une taille d'échantillon donnée est trop petite pour produire des résultats utiles. Cette connaissance est essentiellement dérivée de l'expérience d'enquêtes antérieures (la vôtre et celle des autres) sur des thèmes semblables soumis à des analyses semblables. Dans la pratique, aucune enquête n'est entièrement nouvelle ni menée dans l'isolement. Dans la planification d'une enquête, on commence par se poser des questions telles que celles-ci:

- Quelle taille d'échantillon a-t-on utilisé dans le passé pour cette sorte d'enquête au contexte similaire?
- Quelle sorte d'informations (avec quels détails? avec quelle précision?) ont-elles produit? Les expériences négatives – par exemple lorsqu'une taille d'échantillon s'est révélée inadéquate – peuvent être particulièrement instructives.
- Est-il avéré que les tailles d'échantillon, bien que justifiées sur le plan des conditions de précision, ont été trop importantes au point d'influer

négativement sur la mise en œuvre de l'enquête et la qualité des données obtenues ?

- En quoi les conditions et les exigences actuelles diffèrent-elles de celles de ces enquêtes antérieures ? Par exemple, demande-t-on, dans les résultats, des classifications plus (ou moins) détaillées ?
- Que sait-on déjà sur le sujet à étudier? Quel niveau de connaissances existant la nouvelle enquête doit-elle améliorer ?

Il est particulièrement important de chercher à déterminer *les conditions minimales pour que l'enquête soit suffisamment utile*, étant donné son coût d'opportunité. Comme on l'a indiqué, aucune estimation isolée ne peut en règle générale être considérée comme «cruciale» dans la détermination de ces conditions minimales: mais des types d'estimations différents revêtent une importance relative différente, et il est souvent possible d'identifier un sous-ensemble d'objectifs qui peuvent être considérés comme fondamentaux pour déterminer les conditions minimales de taille d'échantillon.

Contrôler les erreurs d'échantillonnage

Des estimations d'enquête sont normalement exigées non seulement pour la population entière, mais aussi, et séparément, pour de nombreux sous-groupes de la population. L'ampleur relative de l'erreur d'échantillonnage par rapport aux autres types d'erreur augmente lorsque l'on passe d'estimations concernant la population totale à des estimations concernant des sous-groupes distincts et des différences entre sous-groupes. *Au-delà d'un certain niveau de classification détaillée, l'erreur d'échantillonnage prédominerait sur les autres sources d'erreur.*

Le nombre de domaines de rapport principaux

Comme nous l'avons indiqué, la taille d'échantillon minimale requise augmentera avec le nombre de sous-groupes de population ou de domaines pour lesquels les résultats doivent faire l'objet de rapports séparés. Toutefois, pour plusieurs raisons, cette augmentation est en général moins que proportionnelle. On peut examiner, par exemple, les incidences de la nécessité de produire des estimations séparées pour des régions distinctes d'un pays.

- Généralement, on dispose déjà d'informations plus précises au niveau national qu'au niveau de chaque domaine subnational; pour ajouter quelque chose de nouveau à ce qui est déjà connu, l'enquête doit répondre à des conditions de précision plus strictes à l'ensemble du niveau national.
- Les conditions de précision sont également moins strictes pour les domaines distincts que pour le pays tout entier, car les résultats obtenus au niveau national présentent habituellement davantage d'intérêt sur les plans décisionnel et fondamental.
- Des analyses plus variées et détaillées sont nécessaires au niveau global, par comparaison avec celles qui sont requises pour des domaines distincts.
- Ces domaines sont souvent moins hétérogènes que la population nationale dans son ensemble et exigent donc un échantillon plus limité pour obtenir la même précision.

- Pour certains types de domaine, comme les sous-classes bien réparties dans la population, le plan d'échantillonnage est généralement plus efficient (les effets de l'échantillonnage par grappes étant plus limités), si bien qu'on peut obtenir la même précision avec une taille d'échantillon plus réduite.

La limite supérieure dans la pratique

À l'autre extrémité de l'échelle, des considérations pratiques de coût, d'opportunité et de contrôle de qualité déterminent – peut-être avec une acuité et une rigidité moindres – la limite supérieure praticable en matière de taille d'échantillon. L'augmentation de cette taille est d'un rendement décroissant en ce qui concerne la précision d'échantillonnage des résultats. (L'erreur d'échantillonnage n'est réduite approximativement qu'en proportion de la racine carrée de l'augmentation de taille). Quant aux effets négatifs de l'augmentation de la taille d'échantillon sur le contrôle de qualité, ils tendent à s'amplifier – et dans bien des cas à exploser – au-delà d'une certaine taille, cependant que l'enquête devient ingérable.

L'augmentation de la taille d'échantillon, au moins au-delà d'une certaine limite, peut avoir des incidences négatives sur tous les aspects de la qualité des données – pertinence, opportunité, qualité des réponses, etc.

Bien entendu, des considérations de coût peuvent imposer une limite beaucoup moins flexible à la taille d'échantillon maximale admissible.

Le compromis entre les limites minimale et maximale

Dans la pratique, le processus de détermination de la taille d'échantillon peut fonctionner selon les principes suivants. Étant donné les grands domaines pour lesquels des résultats distincts doivent être rapportés, et le type d'estimations requis, la taille d'échantillon minimale peut être déterminée afin de répondre aux conditions de précision les plus cruciales. Il importe toutefois que la taille ainsi déterminée n'excède pas le maximum praticable, compte tenu des contraintes imposées. Si le choix initial excède ce maximum, il vaut mieux reconsidérer et ajuster les objectifs et les conditions d'établissement des rapports de l'enquête, plutôt que d'essayer d'imposer une taille d'échantillon d'une ampleur irréaliste. Quand cette contradiction ne peut pas être résolue, l'annulation de toute l'enquête peut être la seule option.

Lorsqu'il existe une latitude de compromis entre les limites minimale et maximale, le choix de la taille d'échantillon dépend de plusieurs considérations. La valeur de l'enquête augmente avec celle de la taille d'échantillon, puisqu'une analyse plus précise et détaillée de ses résultats devient possible. Par contre, il peut y avoir des incidences défavorables sur la qualité, l'opportunité et le coût de l'enquête. Équilibrer ces considérations contradictoires est une affaire de jugement fondé sur des données théoriques, des informations empiriques, l'expérience passée et la connaissance des conditions pratiques de l'enquête – même si aucun de ces facteurs ne peut en lui-même fournir des résultats quantitatifs précis.

Des conditions différentes pour des enquêtes de types différents

L'expérience a établi des normes utiles pour de nombreux types d'enquête dont nous pouvons tirer des principes directeurs raisonnables. Un point important est à noter: si la théorie et la pratique ne peuvent pas apporter des réponses précises sur la taille des échantillons en termes absolus, elles peuvent apporter une aide beaucoup plus importante si les questions sont posées en *termes relatifs.* Par exemple:

- Pour une enquête sur un thème donné, la taille d'échantillon devrait-elle être plus grande, plus petite ou identique par comparaison avec les enquêtes passées menées dans d'autres pays ? Par comparaison avec les normes établies lorsqu'elles existent ? Si oui, dans quelle proportion (même approximative), étant donné les objectifs et les conditions différents des enquêtes ?
- Quelle devrait être la relation, en matière de taille d'échantillon, entre les enquêtes de types différents portant sur des thèmes différents, mais conduites dans des circonstances semblables ?

Sur le second point, nous savons, par exemple, que les enquêtes impliquant des mesures physiques ou des suivis intensifs gagnent à être conduites à l'aide d'échantillons relativement restreints, alors que les enquêtes portant par exemple sur la main-d'œuvre, qui doivent produire des estimations relativement simples mais détaillées, ont besoin d'échantillons beaucoup plus importants. On dispose à ce propos de beaucoup d'informations issues de l'expérience de différents pays et de différents types d'enquêtes.

Dans des circonstances données, des types d'enquêtes différents tendent à présenter moins de diversité en termes d'«échelle» qu'ils représentent qu'en simples termes de taille d'échantillon. Comme on l'a indiqué, il faut entendre par «échelle» la charge totale que constitue une enquête, en fonction de la taille de l'échantillon, de la sensibilité et de la lourdeur de l'information, de la complexité et du volume de l'information collectée par cas traité, et bien entendu du coût unitaire. Nous pouvons nous attendre à ce que les *tailles d'échantillon appropriées* pour des enquêtes plus complexes, difficiles et coûteuses (comme celles qui portent sur les revenus et les budgets des ménages) soient normalement plus petites que pour des enquêtes plus simples et moins onéreuses (comme celles qui portent sur la main-d'œuvre), les différences entre tailles d'échantillon appropriées augmentant avec les différences de complexité et de coût unitaire.

Les enquêtes destinées à évaluer l'ampleur du travail des enfants (ETE) auront normalement une taille d'échantillon plus petite (et souvent beaucoup plus petite) que dans le cas des enquêtes nationales sur la main-d'œuvre, mais beaucoup plus importante que dans le cas des enquêtes désignées pour l'évaluation de certains aspects du travail des enfants (EET).

3.5.2. La spécification de la précision requise

La précision de l'échantillonnage en terme de taille d'échantillon efficace

La précision de l'échantillonnage est déterminée par la taille de l'échantillon, ainsi que par son plan, autrement dit son efficience ou l'«effet du plan». Ces deux facteurs sont spécifiques aux statistiques envisagées.

Supposons que nous avons identifié les statistiques ou l'ensemble des statistiques les plus cruciales pour les objectifs énoncés. Même ainsi, il est utile de séparer la question des caractéristiques de l'effet du plan de la structure d'échantillon autre que la taille d'échantillon. Les conditions de précision sont plus clairement exprimées et comprises en terme de taille d'échantillon «efficace» plutôt que de taille réelle. Par «taille efficace» d'un échantillon à plan complexe, nous entendons la taille d'un échantillon aléatoire simple d'unités d'analyse possédant la même précision que le plan complexe. La taille efficace d'un échantillon complexe de taille n avec un effet de plan deft^2 est

$$n_{eff} = \frac{n}{deft^2}$$

L'effet du plan de sondage (en anglais «design effect»), ou deft^2 (ou sa racine carrée, *deft*, parfois appelée facteur de plan), est une mesure récapitulative globale de l'effet des divers aspects complexes du plan de sondage sur l'erreur d'échantillonnage. Le deft est le ratio d'erreur type (d'une statistique donnée) lié au plan réel de l'échantillonnage par rapport à ce qu'aurait été l'erreur avec un échantillon aléatoire de la même taille. (Ce concept est examiné plus en détail dans un chapitre ultérieur.)

Les niveaux d'erreur relative et absolue

Pour les estimations de moyennes ou de ratios généraux, les conditions de précision sont commodément exprimées en termes d'erreur type relative (r), c'est-à-dire d'erreur type en pourcentage de la valeur moyenne:

$$n_{eff} = \left(\frac{cv}{r}\right)^2$$

cv étant le coefficient de variation de la variable étudiée entre les éléments individuels de la population.

Pour une proportion (p), cv est estimé grâce à une formule très simple, $cv^2=(1-p)/p$. Toutefois, dans des moyennes et des statistiques plus complexes, des informations empiriques sont nécessaires pour estimer la valeur du paramètre cv dans la population. Il est commode que ce paramètre soit hautement «transférable» d'une situation à l'autre et puisse souvent être estimé très aisément et de manière fiable à partir d'enquêtes passées, d'autres sources ou de populations similaires. Par exemple, pour le revenu des ménages, on relève le plus souvent un cv situé dans une fourchette de 0,7 à 1,0 selon la diversité des populations. L'équation ci-dessus implique que dans ce cas, avec cv=0,7-1,0, une erreur relative de 1 pour cent (r) exigerait une taille d'échantillon

efficace de 5 000-10 000. De même, une erreur relative de 2 pour cent exigerait une taille d'échantillon efficace de 1 250-2 500. La taille d'échantillon réelle devra être plus importante, dans la mesure où deft[2] dépasse 1,0 pour le plan spécifique choisi.

Note sur la terminologie

Il est très courant, dans les rapports d'enquête, d'utiliser les termes «erreur» ou «niveau d'erreur», qui combinent le concept d'erreur relative défini ci-dessus (erreur type en pourcentage de l'estimation à laquelle elle se réfère) avec celui de niveau de confiance[9]. Cette pratique est basée sur l'attente du fait qu'il serait plus utile et commode, pour le lecteur non spécialisé, de connaître avec un certain «niveau de confiance» la fourchette dans laquelle se situe la vraie valeur de la statistique examinée. S'il est possible de choisir des niveaux de confiance différents à des fins différentes, un critère couramment admis établit la valeur de population estimée d'après l'échantillon dans une fourchette au double de l'erreur type de chaque côté de l'estimation. Cela peut être affirmé avec un niveau de confiance élevé (95 pour cent); autrement dit, il est avéré qu'il y a seulement une chance sur vingt pour que les vraies valeurs se trouvent en dehors de cette fourchette (il s'agit là d'un résultat théorique basé sur la normalité supposée de la répartition de l'échantillonnage).

Dans cette formulation, l' «erreur» ou le «niveau d'erreur» (appelons-le R_z) est l'erreur (r) multipliée par un facteur (z): $R_z=z\cdot r$. La valeur de z dépend du niveau de confiance souhaité. Dans l'exemple ci-dessus, z=2 pour un niveau de confiance de 95 pour cent. Des valeurs de z plus élevées donnent des niveaux de confiance supérieurs. Par exemple, avec z=3, on peut dire que les chances sont approximativement d'une sur cent pour que la vraie valeur se situe en dehors d'une fourchette, triple de l'erreur type de chaque côté de l'estimation.

En termes de R_z, la formule ci-dessus pour n_{eff} est

$$n_{eff}=z^2\bullet(cv/R_z)^2$$

Le choix approprié de z est spécifique aux objectifs et aux implications des conclusions à tirer de l'enquête, ce qui dépend de la nature et du contexte de l'information envisagée. Il implique aussi un certain degré d'arbitraire ou de simple convention. Pour ces raisons, il nous paraît préférable de spécifier et de présenter l'information sur les erreurs d'échantillonnage en termes d'erreur type, que ce soit en termes absolus (e) ou relatifs (r), le lecteur restant libre de convertir cela en un intervalle du niveau de confiance souhaité.

La spécification pour les proportions

Dans les enquêtes sur le travail des enfants (tout comme pour beaucoup d'autres types d'enquête), la plupart des statistiques intéressantes revêtent probablement la forme de proportions plutôt que de moyennes ou de ratios généraux. Pour les proportions ou les pourcentages, il importe de distinguer clairement l'erreur exprimée en termes relatifs (en

[9] En fait, les termes mentionnés ci-dessus sont souvent utilisés pour faire référence aux multiples *d'erreurs relatives* pour les valeurs moyennes, mais d'erreurs en *points de pourcentage absolu* pour les proportions.

pourcentage de la proportion p) de celle en termes de points de pourcentage absolus[10]. Les deux formes sont pertinentes. Pour les proportions élevées, il est souvent préférable d'exprimer l'erreur en termes relatifs, alors que l'expression en termes de points de pourcentage absolu est souvent plus significative pour de très faibles proportions [11].

La taille d'échantillon efficace en termes de précision requise pour une proportion (p) peut s'écrire ainsi:

$$n_{eff} = \left(\frac{p*(1-p)}{e_p^2} \right) = \left(\frac{(1-p)}{p*r_p^2} \right)$$

e_p étant l'erreur type en points de pourcentage absolus, et r_p=(e/p) cette erreur exprimée en termes relatifs. L'indice «p» adjoint à ces quantités vient souligner que, dans des situations pratiques, les conditions de précision réalistes dépendent de la valeur de la proportion p considérée. En fait, si p diminue, la valeur significative de e_p est également réduite, tandis que celle de r_p tend à augmenter dans les mêmes proportions. Par exemple, si e=5 % est un choix raisonnable pour l'estimation d'une proportion d'environ p=0,50, alors le même choix du niveau d'erreur e n'est pas rationnel, au moins pour les très petites valeurs de p. De même, si une erreur relative de r=25 %, par exemple, peut être acceptable dans le cas d'une très faible proportion, elle peut être inacceptable pour une proportion élevée telle que p=0,50. Ce point important est souvent oublié dans les débats sur les conditions de précision pour les différentes valeurs de p rencontrées dans une enquête.

En fait, il n'est pas déraisonnable - sauf s'il existe d'autre raisons d'envisager des valeurs différentes de ce terme pour des estimations différentes - de prendre, au moins en tant que point de départ, n_{eff} comme constante pour les différentes proportions p à estimer dans une enquête donnée. Cela impliquerait de prendre e_p comme proportionnel à la racine carrée de [(1-p)*p] pour la série de valeurs de la proportion p rencontrée. Par exemple, si une erreur absolue e=5 % s'adresse à une proportion p=0,5, la valeur appropriée correspondante sera inférieure, soit e=4 %, pour une proportion p=0,2 (ou 0,8); pour p=0,1 (ou 0,9), nous aurons e=3 %, se réduisant à e=1 % pour p= 0,01 (ou 0,99). Cela représente assurément une variation plus significative que de supposer une valeur constante pour e (ex.: e=5 % même quand p n'est que de 0,1, voire même 0,01). Si la supposition précise évoquée ci-dessus peut être contestée, la variation impliquée dans les conditions de précision va du moins dans la bonne direction.

Un modèle semblable mais opposé prévaut en ce qui concerne la précision requise exprimée en termes relatifs. L'hypothèse consiste à prendre r_p comme proportionnel à la racine carrée de [(1-p)/p] pour la série de valeurs de la proportion p rencontrée. Dans notre exemple, nous avons r=10 % pour p=0,5, 20 % pour p=0,2, 30 % pour p=0,1, et un accroissement de r jusqu'à 100 % pour une valeur p=0,01 extrêmement faible. Nous l'avons dit, ces chiffres apparaissent plus significatifs que l'hypothèse d'une

[10] Exemple: un taux de travail des enfants de 22 % diffère d'un taux de 20 % de 10 % en termes relatifs, mais de 2 points de pourcentage en termes absolus.

[11] Cela est exact sur le plan de la conception statistique. Il est toutefois courant, dans les présentations destinées au grand public, de citer les erreurs en termes de points de pourcentage absolu (généralement +/- 2 fois les erreurs type). Pour les moyennes, il est plus significatif de rapporter les erreurs en termes relatifs, car l'échelle de mesure est essentiellement arbitraire dans ce contexte.

exigence de niveau constant d'erreur relative, abstraction faite de la valeur de p concernée.

On notera que l'hypothèse particulière de l'illustration ci-dessus implique que les conditions de précision varient en fonction de p de telle sorte que la taille d'échantillon efficace qui en résulte est indépendante de la valeur p considérée. Cette forme de variation, selon l'hypothèse ci-dessus, peut en fait être simpliste: elle n'est censée être qu'une illustration, adoptée pour son extrême simplicité. Elle n'en fournit pas moins un point de départ utile pour la spécification des conditions de précision e_p et r_p (respectivement en termes absolus et relatifs) en tant que fonctions de la proportion p considérée.

Dans la pratique, une augmentation de n_{eff} avec p en diminution est normalement assurée.

En bref, il est recommandé que, dans l'examen de la taille d'échantillon et des conditions de précision pour les proportions, l'erreur d'échantillonnage soit prise en considération en termes aussi bien absolus que relatifs. Ces conditions de précision doivent varier en fonction des différentes proportions (p) à estimer d'après l'enquête, mais de manière à limiter fortement la variation impliquée avec p de la taille d'échantillon requise.

Le tableau 3.1 illustre la manière dont la taille d'échantillon peut être accrue avec une proportion (p) décroissante à estimer, en tant que fonction de p. Nous supposons que la taille d'échantillon a varié selon

$$n \propto 1/p^a$$

où a est un paramètre, qui figure dans la fourchette 0-0,25 sur le tableau. On a pris la base de n=100 pour p=50 %. Prenons le cas de a=0,10, qui peut en fait constituer un choix raisonnable dans la pratique. Cela implique d'augmenter la taille d'échantillon, par exemple, de 100 à p=50 % jusqu'à 120 pour une estimation de proportion réduite p=20 %, à environ 160 pour p=5 %, et à 220 pour un p=1 % très faible. On notera la variation impliquée dans les niveaux d'erreur relatifs et absolus requis avec variation de p. Dans la pratique, une taille d'échantillon presque constante peut être acceptable pour un large éventail de valeurs p – par exemple n=120 pour p=10-30 % dans cette illustration.

Tableau 3.1: Quelques illustrations de la variation de la taille d'échantillon selon la proportion estimée

	a=0			0,05			0,10			0,20			0,25		
p(%)	n	%e(p)	%r(p)	n	%e(p)	%r(p)	n	%e(p)	%r(p)	n	%e(p)	%r(p)	n	%e(p)	%r(p)
50,0	100	5,0	10	100	5,0	10	100	5,0	10	100	5,0	10	100	5,0	10
40,0	100	4,9	12	102	4,8	12	105	4,8	12	109	4,7	12	112	4,6	12
30,0	100	4,6	15	105	4,5	15	111	4,4	15	123	4,1	14	129	4,0	13
20,0	100	4,0	20	110	3,8	19	120	3,6	18	144	3,3	17	158	3,2	16
10,0	100	3,0	30	117	2,8	28	138	2,6	26	190	2,2	22	224	2,0	20
5,0	100	2,2	44	126	1,9	39	158	1,7	35	251	1,4	28	316	1,2	25
2,0	100	1,4	70	138	1,2	60	190	1,0	51	362	0,7	37	500	0,6	31
1,0	100	1,0	99	148	0,8	82	219	0,7	67	478	0,5	46	707	0,4	37

p%= Pourcentage à estimer
a= Paramétre determinant n
n croissant avec p décroissant p: n=k/(p ^ 2a)
n= Taille d'échantillon (effective)
%e(p)= Erreur en points de pourcentage (absolus)
%r(p)= Erreur relative (%)

3.5.3. Les facteurs qui déterminent la taille réelle de l'échantillon

La variation de l'effet du plan de sondage avec la taille des grappes

Dans un plan à plusieurs degrés, la taille d'échantillon requise pour atteindre un niveau de précision donné dépend de l'effet du plan de sondage. L'effet de l'échantillonnage par grappes peut être résumé en deux paramètres: la corrélation intragrappes (roh) et la taille des grappes (b)[12].

Le premier paramètre (roh) est une mesure de l'homogénéité relative des éléments provenant des mêmes grappes d'échantillons. Il dépend avant tout de la nature des variables d'intérêts (les valeurs rencontrées pour des variables différentes peuvent différer largement). La nature et la structure de l'échantillonnage au sein des grappes d'échantillons ont aussi un effet.

Le second paramètre (b) est le nombre moyen d'éléments (les unités finales concernées par l'analyse) sélectionnés par UEP dans l'échantillon. On peut l'exprimer, pour une proportion p par exemple, sous la forme suivante:

$$n = n_{eff} \cdot deft^2 = \left[\frac{p*(1-p)}{e_p^2}\right][1+(b-1)roh]$$

Sur la base des calculs d'erreurs d'échantillonnage provenant d'enquêtes passées comparables, on peut se former une idée raisonnable du modèle de valeurs roh rencontré pour différents types de variables et pour divers plans.

Pour un niveau de précision requis (exprimé en termes de e_p ou n_{eff}), le choix de la taille b des grappes (et par suite le nombre des grappes à sélectionner, a=n/b) affecte la taille d'échantillon requise n. Le tableau 3.2 présente le facteur $(n/n_{eff})=(1+(b-1).roh$ par

[12] Les erreurs d'échantillonnage sont examinées plus en détail dans le chapitre 7.

lequel la taille d'échantillon requise doit être augmentée, en fonction des valeurs roh et b pour la variable et le plan d'échantillon concernés.

Tableau 3.2: L'augmentation requise pour la taille d'échantillon en fonction de l'augmentation de la taille des grappes (b) et de la corrélation intragrappes (roh)

		roh				
		0,00	**0,02**	**0,05**	**0,10**	**0,20**
b	1	1,00	1,00	1,00	1,00	1,00
	5	1,00	1,08	1,20	1,40	1,80
	10	1,00	1,18	1,45	1,90	2,80
	20	1,00	1,38	1,95	2,90	4,80
	50	1,00	1,98	3,45	5,90	10,80

Le coefficient de corrélation intragrappes (roh) est une mesure plus «transférable» que le deft en ce sens qu'il abstrait également l'effet de la taille moyenne de grappe b. Il est défini en termes de deft et de b comme

$$roh = \frac{deft^2 - 1}{(b-1)}$$

Le point pratique à souligner est que la taille d'échantillon n requise pour obtenir un degré spécifié de précision des estimations (ou n_{eff} spécifié) ne peut pas être déterminée indépendamment de l'efficience ou $deft^2$ du plan. Dans un échantillon en grappe, les deux paramètres, à savoir *la taille d'échantillon et le nombre de zones de l'échantillon* (ou encore le prélèvement moyen d'échantillons par zone ou par grappes) *doivent être déterminés simultanément*. Plus la corrélation intragrappes (roh) est importante, plus fortes seront les relations.

Les effets des variations dans les poids de l'échantillon

L'enquête est souvent conçue pour produire des types d'estimations différents à des niveaux d'agrégation différents. Les différents objectifs suscitent des conflits, qui exigent des compromis dans la répartition de l'échantillon. Un exemple courant est celui des plus petits domaines, qu'il est nécessaire de suréchantillonner pour obtenir une taille d'échantillon plus adéquate, et par suite sous-échantillonner les domaines plus importants de manière correspondante. Pour tout objectif donné, comme la production d'estimations pour l'ensemble de la population, la répartition de compromis représente une pondération qui est par essence aléatoire ou arbitraire.

Cette pondération a un effet inflationniste sur les variances concernant cet objectif. Cet accroissement tend à être assez uniforme pour les différents types de variables et de groupes de population du domaine concerné. L'augmentation de la variance est donnée approximativement par $[1+cv^2(w)]$ en terme de coefficient de variation des pondérations de chaque échantillon.

Unités d'analyse versus unités d'échantillonnage

Bien entendu, les conditions de précision et de taille d'échantillon sont spécifiées en fonction des unités qui intéressent l'analyse. Il peut ainsi s'agir de catégories particulières d'individus, comme les adultes, les personnes âgées, les femmes ou les enfants. Celles-ci peuvent être différentes des unités finales utilisées dans l'échantillonnage, qui peuvent par exemple être des adresses ou de ménages. Sur le plan opérationnel, la taille d'échantillon doit être exprimée en fonction de ces derniers types d'unités.

Supposons que le ménage soit l'unité finale d'échantillonnage et qu'un ménage contienne en moyenne h unités d'analyse d'intérêt (ex.: h adultes, ou h enfants travailleurs). La taille d'échantillon requise en terme de nombre de ménages (n_h) peut s'exprimer ainsi:

$$n_h = \frac{n}{h} = \left(\frac{n_{eff}}{h}\right) \cdot \left[1 + (h * b_h - 1).roh\right]$$

Ici, b_h est le nombre moyen de ménages par grappe d'échantillons et, comme précédemment, n_{eff} est la taille d'échantillon efficace requise aux termes des *unités d'analyse*. En voici quelques exemples:

1. La situation la plus simple est celle dans laquelle les deux unités sont du même type: h=1. Nous pouvons la prendre comme critère de comparaison.

2. Supposons que dans un échantillon de ménages, les unités d'analyse soient les enfants en âge d'aller à l'école de 6 à 14 ans. Le nombre moyen de ces enfants par ménage sera probablement largement inférieur à 1,0; disons que h=1/3. Nous devons admettre proportionnellement davantage d'enfants dans l'échantillon pour obtenir une taille d'échantillon de n enfants. En même temps, la taille efficace des grappes est proportionnellement réduite à $(h*b_h)=b_h/3$. Nous pouvons donc augmenter le nombre des ménages par grappe lorsque h diminue sans augmenter nécessairement l'effet du plan de sondage pour une valeur de roh donnée.

3. Supposons que nous nous intéressons à un échantillon d'adultes et que nous intégrons dans l'échantillon tous les adultes de chaque ménage sélectionné. Le nombre des unités d'analyse par unité d'échantillonnage sera supérieur à 1,0; disons que h=2. Pour obtenir un échantillon de n adultes, nous aurons proportionnellement besoin de moins de ménages. Pour maintenir la même efficience, nous devons réduire le nombre de ménages de l'échantillon par grappe afin de préserver le même effet du plan de sondage pour une valeur de roh donnée.

4. Une solution de substitution au plan (3) consiste à sélectionner exactement un adulte pour chaque ménage de l'échantillon. Dans la mesure où chaque ménage contient au moins un adulte par définition, h=1 et la situation apparaît comme semblable à (1). Il existe toutefois quelques différences importantes entre les deux situations. D'un côté, le plan (4) évite l'homogénéité ajoutée entre personnes provenant du même ménage, réduisant ainsi l'effet du plan de sondage par rapport au plan (1). D'un autre côté, cependant, à supposer que les ménages aient été

choisis avec des probabilités uniformes, les probabilités d'échantillonnage relatives des individus se trouvent altérées et deviennent inversement proportionnelles au nombre de personnes sur lequel le ménage aurait pu être choisi. Cela exige une pondération au stade de l'estimation, qui influe elle-même sur l'efficience de l'échantillon, et par suite sur la taille d'échantillon réelle requise pour un n_{eff} donné.

Il est fréquent que ce second effet négatif prédomine, rendant le plan (4) moins efficient que l'échantillonnage direct des personnes pratiqué en (1).

Les pertes dues aux non-réponses

De surcroît, une partie de l'échantillon est perdue pour cause de non-réponse. La taille de l'échantillon sélectionné doit être plus importante que la taille d'échantillon obtenue grâce au facteur (1/R), dans lequel R est le taux de réponse.

3.5.4. L'expérience des enquêtes sur le travail des enfants

Certaines informations sur la taille des échantillons et les structures d'enquêtes antérieures sur le travail des enfants ont été données dans le tableau 2.1.

À partir des enquêtes sur le travail des enfants que nous avons étudiées, nous avons dressé le tableau général suivant. Voir la section 2.1.5 pour des informations complémentaires.

Une grande majorité des tailles d'échantillon de la composante ETE se situent dans la fourchette 8.000-20.000, bien que des tailles beaucoup plus importantes aient été utilisées dans quelques cas. On dispose de peu d'informations dans les publications et les rapports nationaux sur la pertinence de ces choix par rapport à des objectifs spécifiques dans des cas particuliers. Nous avons malheureusement constaté que la même situation prévaut pour beaucoup d'autres types d'enquêtes. Il est nécessaire de déterminer dans quelle mesure ces choix de taille d'échantillon ont été appropriés, afin d'apporter une orientation pour l'amélioration des pratiques dans l'avenir.

3.5.5. Conseil pratique: de la modération dans le choix de l'échantillon

La taille d'échantillon doit, bien entendu, être assez importante pour répondre aux conditions fondamentales et spécifiques de l'enquête. Il n'en reste pas moins que, dans un grand nombre de cas, *la qualité des enquêtes a souffert du choix de taille d'échantillon d'une importance inappropriée*. Ces choix inappropriés peuvent résulter de plusieurs facteurs:

- valorisation excessive de la précision de l'échantillonnage et négligence de la nécessité de contrôler les erreurs non dues à l'échantillonnage et d'assurer la pertinence et l'opportunité des résultats;
- adoption de normes de précision déraisonnables et inutilement élevées;
- désir d'établir trop de répartitions à un niveau de détail excessif;
- défaut d'exploration de procédures de substitution visant à renforcer la précision des résultats, comme la cumulation des résultats dans le temps grâce

à des enquêtes périodiques, utilisation de modèles et de relations appropriés, ajustement des données d'enquête sur la base de totalisations de contrôle provenant de sources extérieures plus fiables, etc.;

- utilisation de plans inefficients, qui augmentent la taille d'échantillon requise pour atteindre un niveau de précision donné; et par-dessus tout,
- sous-estimation du coût unitaire et des efforts requis dans la collecte, le traitement et l'analyse des données.

3.6. L'échantillonnage des unités de zone avec probabilité proportionnelle à la taille (PPT)

Il s'agit là, et de loin, de la technique la plus courante pour la sélection d'échantillons basés sur la population, notamment dans les pays en développement. Le plan implique la sélection d'unités de zone selon un processus à un ou plusieurs degrés (souvent un seul) *avec probabilité proportionnelle à la mesure de la taille de la population* de la zone (p_i) et, au sein de chaque zone sélectionnée, la sélection d'unités finales *avec probabilité inversement proportionnelle à la mesure de la taille:*

Sélection des zones

$$f1_i = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}, \text{ disons.} \qquad [1]$$

Sélection des unités finales au sein de la zone sélectionnée

$$f2_i = \left(\frac{b}{p_i}\right) \qquad [2]$$

Probabilité globale de sélection d'une unité finale

$$f_i = f1_i * f2_i = \left(\frac{b}{I}\right) = f, \text{ une constante.} \qquad [3]$$

La sommation Σ porte sur toutes les zones de la population. Ici, a renvoie au nombre de zones (au sens strict des «unités finales de zone», UFZ) sélectionnées. Si p_i est exactement le nombre actuel d'individus concernés dans chaque zone, alors b=constant est le nombre d'unités finales sélectionnées dans n'importe quelle zone de sondage, et n=a*b est la taille d'échantillon résultante.

Dans le cas idéal, un tel plan présente donc le double avantage: (i) d'un contrôle de la taille d'échantillon et d'une charge de travail fixe b par zone de sondage; et (ii) d'une probabilité d'échantillonnage globale uniforme f pour chaque unité finale (ex.: chaque ménage, enfant ou autre personne).

Dans la pratique, il est peu probable que ces deux conditions – et même l'une d'elles – soient exactement remplies. D'une part, l'unité en fonction de laquelle les mesures de

taille sont disponibles est souvent différente du type d'unité utilisée au stade ultime de la sélection, qui peut lui-même être différent des unités d'analyse; il peut s'agir, par exemple, de personnes par rapport à des ménages ou d'habitations par rapport à des enfants. Mais le vrai problème est que dans la plupart des circonstances, les mesures de taille disponibles dans la base sont obsolètes et peuvent ne pas concorder avec les tailles réelles (actuelles) des unités de zone.

On trouvera ci-dessous (section 3.8) la description de quelques variantes courantes (et souvent proches) de ce plan de base. Il est utile de clarifier d'abord la pratique d'une méthode de sélection d'échantillons très couramment utilisée, à savoir la procédure de tirage systématique.

3.7. L'échantillonnage systématique

Une méthode courante d'échantillonnage consiste à sélectionner systématiquement les unités sur une liste ordonnée. L'échantillonnage systématique avec probabilité égale est couramment utilisé pour la sélection des unités finales, telles que les ménages ou les personnes, dans les zones de sondage. Plus importante est l'utilisation de l'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT), communément pratiqué pour la sélection des unités de zone.

On trouvera ci-dessous une brève description formelle de la procédure; pour une illustration chiffrée, cf. la section 6.7.

3.7.1. L'échantillonnage systématique avec probabilité égale

La procédure de sélection systématique sur une liste d'un échantillon à probabilité égale se déroule essentiellement comme suit. Supposons qu'un échantillon à probabilité égale de n unités (entrées sur une liste) doit être obtenu à partir d'une population de N unités. Sur la liste de ces unités - de préférence ordonnées d'une manière utile avec une numérotation séquentielle de 1 à N -, on choisit une seule unité sur chacune des $I=N/n$ unités de la liste.

En premier lieu, supposons pour simplifier que l'intervalle de sélection I se trouve être un nombre entier. Un nombre aléatoire r entre 1 et I identifie le numéro d'ordre de la première unité sélectionnée. En commençant par r, on sélectionnera ensuite chaque I^e unité. Autrement dit, la séquence des numéros sélectionnés sera:

$$r, r+I, r+2*I, \ldots\ldots, r+(n-1)*I$$

Il est facile de traiter le cas général, dans lequel $I=N/n$ n'est pas un nombre entier. Une des procédures utilisables est la suivante. En commençant par un nombre aléatoire réel (et non pas nécessairement un nombre entier) dans la série $0<r\leq I$, on construit la séquence définie ci-dessus. Chaque terme de cette séquence, arrondi au nombre entier le plus proche, identifie une unité sélectionnée.

Dans la mesure où les unités de la liste originale apparaissent dans un ordre aléatoire, l'échantillon résultant est équivalent à un échantillon aléatoire simple. Toutefois, il est

très rare que les listes existantes soient ordonnées de manière aléatoire. Dans tous les cas, l'échantillonnage systématique vise à faire usage de l'ordre disponible pour réaliser une meilleure répartition de l'échantillon en fonction de quelques critères significatifs, telle la localisation géographique des unités. De cette manière, l'échantillonnage systématique fournit une stratification implicite; il peut être considéré comme une stratification de la population en zones ou strates de taille *I*, et la sélection d'une seule unité par zone ou «strate implicite».

L'utilisation répandue de l'échantillonnage systématique s'explique aussi par sa grande commodité dans de multiples situations.

3.7.2. L'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT)

Pour sélectionner des unités avec probabilité proportionnelle à une mesure de la taille unitaire, telle que la taille de la population ou le nombre d'enfants dans la zone, la principale différence avec ce qui précède est que le simple dénombrement des unités est remplacé par le *cumul de leurs mesures de taille*. Si nous avons

p_i comme mesure de la taille de l'unité i,

P_i comme le cumul de ces mesures pour toutes les unités 1 à i, ordonnées de manière significative,

P comme cette somme pour toutes les unités de la population, et

a comme nombre d'unités à sélectionner avec PPT,

l'intervalle d'échantillonnage à appliquer aux mesures de taille cumulatives est:

$$I=P/a$$

Un nombre aléatoire r dans l'intervalle $0<r\leq I$ identifie la première unité sélectionnée: c'est la première unité dont la taille cumulative est égale ou supérieure à r; autrement dit, le numéro de séquence de l'unité satisfaisant à la relation:

$$P_{i-1} < r \leq P_i$$

Ensuite, en commençant avec r, le point de sélection subit à chaque fois un accroissement de I, ce qui donne une série de ce type:

$$r'=r+I,\ r+2*I,\ .....\ ,\ r+(n-1)*I$$

et chaque unité à mesure de taille cumulative satisfaisant la relation

$$P_{i-1} < r' \leq P_i$$

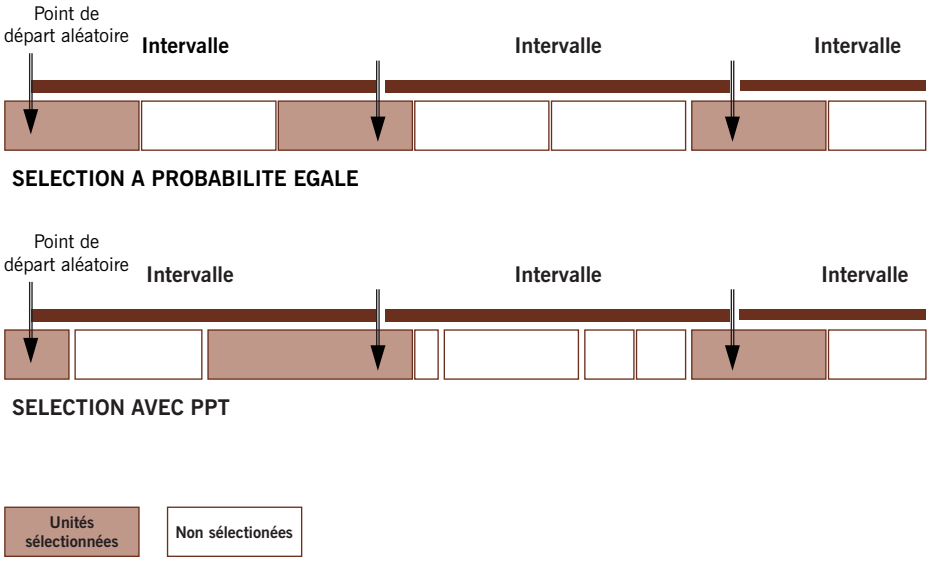
pour tout i dans la liste de mesures cumulative de la taille $P_{i,}$ est sélectionnée. La chance de sélection d'une unité particulière est proportionnelle à

$$P_i - P_{i-1} = p_i$$

c'est-à-dire à la mesure de la taille de l'unité. La valeur absolue de la probabilité de sélection est:

$$f1_i = \left(\frac{a}{P}\right).p_i = \frac{p_i}{I}$$

Figure 1: Illustration de la procédure d'échantillonnage systématique



Le diagramme ci-dessus présente une illustration physique de la procédure d'échantillonnage systématique avec PPT.

3.8. Les mesures de taille imparfaites

Les mesures de taille (p_i) qui déterminent les probabilités de sélection des unités de zone doivent être basées sur des informations antérieures, généralement obtenues au moment de la compilation de la base de zone. En effet, ces mesures sont requises pour toutes les unités de zone de la population et ne peuvent pas être actualisées dans l'échantillon seulement. Elles peuvent donc différer des tailles réelles ou actuelles (p'_i).

L'inexactitude des mesures de taille est un problème courant et par conséquent important. Avec les unités de degré supérieur sélectionnées avec probabilités proportionnelles aux tailles estimées qui s'écartent sensiblement des tailles unitaires réelles, il importe de respecter la distinction entre deux types de plan pour le stade suivant de la sélection:

a) d'une part, un plan strictement auto pondéré des unités finales, et

b) d'autre part, un plan basé sur des tirages fixes.

c) dans le cas des enquêtes sur le travail des enfants, une troisième option peut également être appropriée: l'échantillonnage recensement, éventuellement avec arrêt sur un *maximum de tirage pour toutes les zones.*

3.8.1. Autopondération et plan à tirage fixe

Dans le *plan autopondéré* décrit ci-dessus, les unités finales reçoivent toutes la même probabilité globale de sélection, et le nombre de ces unités sélectionnées peut varier dans la mesure où les mesures de taille utilisées dans la sélection des UEP diffèrent de leurs tailles réelles. D'après l'équation [2], le nombre des unités sélectionnées dans une zone de sondage est:

$$b_i = f2_i * p'_i = b * \left(\frac{p'_i}{p_i} \right) \qquad [4]$$

avec également une variation (toutefois plus faible) dans la taille de l'échantillon d'ensemble $n' = \Sigma_s b_i$. . (Ici, la somme porte sur toutes les autres zones de l'échantillon, comme l'indique l'indice s.[13])

Dans le plan à tirage fixe, le nombre des unités finales par zone de sondage est déterminé par

$$f2_i = \left(\frac{b}{p'_i} \right); \text{nous avons } b_i = f2_i * p'_i = b, \text{ une constante.} \qquad [5]$$

En conséquence, cependant, les probabilités globales d'échantillonnage varient proportionnellement aux inexactitudes des mesures de taille:

$$f_i = f1_i * f2_i = f * \left(\frac{p_i}{p'_i} \right) \qquad [6]$$

Il est courant de tirer des échantillons de taille fixe prédéterminée dans des situations où des listes complètes d'unités ne sont pas disponibles avant la sélection – mais alors le résultat n'est pas, au sens strict, un échantillon probabiliste. On peut préférer les plans à tirage fixe dans les enquêtes très «lourdes» (comportant de très longs entretiens, des visites fréquentes et répétées dans chaque ménage, ou des mesures physiques complexes), quand on estime que les variations, même mineures, du nombre de cas échantillonnés provoquent des variations inacceptables dans la charge de travail des enquêteurs. Mais il s'agit là d'exceptions.

Il semble néanmoins que l'utilisation de plan à tirage fixe plutôt qu'autopondéré soit une pratique très courante; il en va ainsi, par exemple, dans la majorité des enquêtes sur le travail des enfants menées jusqu'ici. Cela est peut-être dû au fait que l'éventuel manque de contrôle sur la taille de l'échantillon global est sans aucun doute l'inconvénient le plus sérieux du plan autopondéré. Des variations excessives de la charge de travail entre

[13] L'autopondération fait référence à la variation (ou à son absence) dans les probabilités de sélection finale entre grappes dans une strate donnée. Elle ne concerne pas les variations possibles dans les taux d'échantillonnage entre strates ou entre domaines. Le même contexte s'applique aux plans à tirage fixe qui sont examinés.

les zones de sondage peuvent aussi représenter un sérieux inconvénient. (Voir dans la section 3.8.3 les procédures de compromis ou de substitution qui permettent d'atténuer le problème.)

Malgré tout, nous recommandons cependant d'opter, en règle générale, pour un plan autopondéré plutôt que pour un tirage fixe. En effet, le tirage fixe présente un certain nombre d'inconvénients pratiques, qui résultent essentiellement du fait que l'on fixe la taille de l'échantillon à obtenir plutôt que d'appliquer les probabilités de sélection.

- En premier lieu, les variations arbitraires dans la sélection finale des probabilités, qui sont inhérentes au plan à tirage fixe, sont généralement indésirables.
- Les probabilités de sélection globales n'étant pas fixées à l'avance, elles ne peuvent être calculées que si l'on comptabilise le nombre réel des unités admissibles listées dans chaque zone de sondage. Cela peut paraître simple, mais de nombreuses enquêtes n'enregistrent pas ces données dans la pratique. Le calcul des pondérations exige que l'on connaisse non seulement le nombre total des unités de la liste, mais aussi celui des éléments admissibles pour l'enquête, par exemple non seulement le nombre des habitations, mais aussi celui des habitations occupées par des ménages privés. Cette condition n'est pas toujours insignifiante et peut exiger la collecte d'informations complémentaires au stade du listage. Cela peut accroître encore la charge du travail de listage, et éventuellement avoir aussi des incidences négatives sur l'exhaustivité de sa couverture.
- Dans la pratique, il est extrêmement courant de sélectionner systématiquement les unités finales à partir de listes. Avec un plan autopondéré, l'intervalle de sélection peut être spécifié préalablement pour chaque zone de sondage. L'intervalle de sélection à appliquer avec un plan autopondéré est basé sur les mesures de taille des zones déjà disponibles dans la base de sondage et ne dépend donc pas du relistage des zones de sondage. De plus, en cas de besoin, les intervalles de sélection spécifiés préalablement peuvent être appliqués à toutes les unités listées, sans distinction entre celles qui sont admissibles et celles qui ne le sont pas, sans pour autant affecter les probabilités de sélection des unités admissibles.
- Cette procédure de sélection est toutefois plus complexe avec un plan à tirage fixe, car l'intervalle de sélection doit être calculé selon le nombre réel d'unités dénombrées dans la zone. Cela peut susciter des problèmes, notamment quand la sélection doit être effectuée par du personnel subalterne sur le terrain.
- La sous-couverture dans les bases de sondage est un problème courant et souvent sérieux, quelle que soit la méthode de sélection de l'échantillon choisie. Toutefois, le problème supplémentaire suivant peut apparaître avec un plan à tirage fixe.
- La comparaison des tailles d'échantillon réellement sélectionnées avec celles qui avaient été ciblées permet de déceler des problèmes de sous-couverture dans la base de sondage, comme le caractère incomplet de la liste des zones incluses dans cette base, ou des erreurs majeures dans la sélection de l'échantillon. Parmi les exemples de ce dernier cas figurent l'omission d'une

partie de la base, les préférences systématiques des agents de terrain pour la sélection de segments réduits et, ce qui est peut-être le plus courant, l'omission systématique due à l'obsolescence ou au mauvais listage des unités finales des zones de sondage. De nombreuses enquêtes ont rencontré le problème d'une taille d'échantillon réellement sélectionnée qui était très éloignée de la taille ciblée ou du plan de l'enquête. Un des inconvénients majeurs d'un plan à tirage fixe est de dissimuler ce type de problème en fixant automatiquement au préalable la taille de l'échantillon.

- Pis encore, un plan à tirage fixe peut réellement encourager les listages incomplets, notamment quand les mêmes agents de terrain sont responsables à la fois du listage et des entretiens. Bien entendu, une supervision plus étroite peut aider à maîtriser ce problème, mais cela sera plus difficile si le listage et le dénombrement sont effectués en une seule opération, avec une chance limitée ou inexistante d'intervention de supervision entre les deux. Il y a eu des enquêtes – y compris au niveau national – pour lesquelles seul un nombre déterminé d'unités, au pire seulement égal à celui qui était requis pour obtenir le nombre fixé de cas échantillonnés, était listé dans chaque zone, ce qui dévalorisait le plan et laissait subsister un énorme potentiel de biais au détriment des unités moins accessibles.
- Fixer la taille des échantillons peut encourager des substitutions incontrôlées pour les cas de non-réponse et décourager les efforts persistants pour les cas les plus difficiles. Cela se produit quand les résultats des enquêteurs sont évalués avant tout sur la base de l'obtention réussie du nombre prédéterminé requis d'entretiens complets dans chaque zone de sondage, plutôt que d'un taux de réponses maximal.
- Même avec un taux fixe d'entretiens pour chaque grappe, la charge de travail réelle demeura variable à cause des inévitables variations dans les circonstances et de la somme d'efforts requise pour chaque entretien dans les différentes zones de l'échantillon.
- Quoi qu'il en soit, il est souvent impossible de fixer les tailles d'échantillon de manière rigide à cause des non-réponses et d'autres problèmes rencontrés au stade de la mise en œuvre. Il en va de même quand les unités à dénombrer ne sont pas du même type que les unités utilisées pour le listing et l'échantillonnage, ou quand l'enquête porte sur une sous-population non contrôlée dans le plan.

Ces problèmes peuvent être encore plus sérieux dans les enquêtes sur les enfants travailleurs, par comparaison avec des enquêtes à base plus large sur la population générale. Le conseil généralement donné est donc d'*adopter, chaque fois que cela est possible, un plan fondé sur l'autopondération* (ou plus généralement et de manière plus appropriée, un plan comportant des probabilités prédéterminées plutôt que des tailles prédéterminées), *et de tolérer les variations dans les tirages par grappe et dans la taille de l'échantillon d'ensemble*. Nous sous-estimons souvent la flexibilité avec laquelle il est possible, dans la pratique, de traiter les variations dans les tirages. Du personnel supplémentaire peut souvent être affecté aux zones où la charge de travail est

exceptionnellement importante. Par ailleurs, il y a davantage de flexibilité lorsque les enquêteurs travaillent en équipes, plutôt que seuls au sein des zones de sondage.

Il y a souvent de bonnes raisons de varier les taux d'échantillon entre les différentes parties de la population enquêtée, telles que le suréchantillonnage des zones urbaines, des petites régions ou des petits domaines présentant un intérêt particulier. Les remarques ci-dessus ne sont pas liées à ce type d'écarts nécessaires par rapport à un plan autopondéré, mais à la nécessité d'éviter les variations inutiles dans les probabilités de sélection des unités finales dans un plan d'échantillonage à degrés multiples.

3.8.2. Le contrôle de la taille de l'échantillon total

Il convient de noter que même dans un plan autopondéré au sens exposé ci-dessus, lorsque la charge de travail individuelle varie considérablement, la taille de l'échantillon total est beaucoup mieux contrôlée que les tirages effectués dans des zones distinctes, dans la mesure où la variation de ces dernières tend à s'annuler quand elles sont agrégées sur toute la série des zones de sondage.

De plus, il est généralement possible de prendre quelques mesures simples pour réduire la variation de la *taille de l'échantillon total.* Par exemple, il peut être envisageable d'ajuster les taux de sélection du dernier degré $f2_i$ sur la base d'informations extérieures plus exactes sur la taille de la population totale (P, remplaçant la mesure de la taille globale $\sum p_i$ dans la base), même si les mesures de taille p_i pour chacune des zones ne peuvent pas être actualisées:

$$f2_i = \left(\frac{b}{p_i}\right) \cdot \left(\frac{\Sigma p_i}{P}\right)$$

Cet ajustement réduit l'impact des changements dans la taille totale de la population sur la taille de l'échantillon obtenu. Ainsi, si la population s'est agrandie (P plus important), le taux d'échantillonnage du degré final $f2_i$ est réduit dans des proportions correspondantes pour maintenir la taille d'échantillon globale requise.

L'ajustement suivant est réalisable quand la taille de la population P n'est pas connue. Si le listage et la sélection de l'échantillon final sont séparées dans le temps et qu'une intervention soit possible entre les deux opérations, la taille de l'échantillon total attendue peut être ajustée en prenant en compte le résultat du listage (qui fournit de nouvelles mesures de taille p'_i) dans les zones de sondage, en changeant le taux d'échantillonnage du dernier degré en:

$$f2_i = \left(\frac{b}{p_i}\right) \cdot \left(\frac{\Sigma_s p_i}{\Sigma_s p'_i}\right)$$

(On notera que dans ce qui précède, les sommations portent seulement sur les zones de sondage, comme l'indique l'indice s, alors que dans l'équation précédente, la sommation porte sur les zones de la population tout entière.)

Les ajustements du type ci-dessus peuvent être utiles, bien que leur intérêt soit limité, dans la mesure où ils affectent seulement la taille de l'échantillon global mais ne réduisent pas la variation des charges de travail pour chacune des zones. En outre, les probabilités de sélection ne sont pas seulement connues en termes relatifs, mais aussi en termes absolus, bien que cela suffise dans la grande majorité des enquêtes par sondage visant à estimer des ratios ou des statistiques semblables plutôt que des agrégats de population, ou lorsque l'on emploie des estimateurs de ratios qui utilisent les totaux issus de contrôles externes pour estimer ces agrégats.

3.8.3. Le contrôle de la taille des échantillons de zone: les solutions de compromis

Plus difficile est évidemment le problème des larges variations de la taille des échantillons selon les zones de sondage. Quand les discordances sont importantes, il peut être très utile d'utiliser une approche qui assurera un compromis entre les options d'autopondération et de tirage fixe. Par exemple, l'utilisation d'une mesure de taille modifiée du type

$$p''_i = (p_i * p'_i)^{1/2}$$

en $f2_i$ peut éviter les variations extrêmes, que ce soit dans les probabilités de sélection globales ou dans le tirage d'échantillons par grappe. Ici p_i est la mesure de la taille de la zone i dans la base de sondage, telle qu'elle est utilisée pour la sélection des zones, et p'_i est la taille réelle de la zone obtenue après listage des zones intégrées dans l'échantillon. La mesure p''_i, compromis entre les deux éléments ci-dessus, est utilisée pour déterminer le taux de sélection comme suit.

Avec un taux d'échantillonnage pour la sélection de ménages dans la zone

$$f2_i = \left(\frac{b}{p''_i}\right)$$

nous avons, pour le tirage d'échantillons dans la zone,

$$b_i = f2_i * p'_i = b.\left(\frac{p'_i}{p_i}\right)^{1/2}, \text{ et } \quad f_i = f1 * f2_i = f.\left(\frac{p_i}{p'_i}\right)^{1/2}$$

Pour les unités dont la taille a quadruplé, par exemple, les tirages comme les taux d'échantillonnage ne changeraient que selon un facteur de 2, et de 3 dans le cas extrême d'un décuplement des mesures de taille. Les effets statistiques et pratiques des variations des tirages ou des taux d'échantillonnage tendent à être négligeables pour les faibles modifications, faibles pour les modifications modérées, mais de plus en plus sensibles pour les modifications importantes. Ces solutions de compromis peuvent donc renforcer l'efficience.

3.8.4. L'échantillonnage recensement, avec arrêt éventuel

Il existe des situations dans lesquelles il est approprié de retenir toutes les unités finales (ménages) de la zone sélectionnée au dernier stade sectoriel. Nous arrivons au stade de la sélection des ménages. Ce système est appelé «échantillonnage recensement» ou «échantillonnage en grappe compacte». L'exemple évident est celui d'une opération de listage des ménages, qui doit couvrir intégralement toutes les zones de sondage.

Dans le contexte des enquêtes sur le travail des enfants, l'échantillonnage recensement résulte souvent de la nécessité d'obtenir pour l'enquête un nombre suffisant d'enfants travailleurs à partir d'un nombre limité de zones de sondage.

Avec l'échantillonnage recensement, la probabilité de sélection des unités finales est la même que celle des grappes dont elles proviennent. Pour obtenir un échantillon autopondéré des premières, par exemple, les zones doivent être sélectionnées avec des probabilités égales. Quand les tailles des zones varient dans des proportions considérables, la stratification par taille en vue de leur sélection peut devenir importante.

Il est également courant de sélectionner des grappes avec probabilité proportionnelle à une certaine mesure de la taille (PPT). Pour l'opération de listage, on procédera normalement ainsi afin de contrôler les tailles d'échantillon pour l'opération *suivante,* comme une EMO ou une ETE, l'échantillon pouvant alors être sélectionné avec PPT inverse selon la description ci-dessus.

Pour les enquêtes sur les enfants travailleurs, il est généralement nécessaire de concentrer l'échantillon dans les zones à concentration plus élevée de ce type d'enfants. Cela implique d'échantillonner la zone avec des mesures de taille définies en fonction du nombre d'enfants travailleurs attendus dans la zone, puis de tirer la totalité ou la majeure partie des ménages (ou des ménages avec enfants travailleurs) de chaque zone sélectionnée. Les probabilités de sélection des unités finales varient en fonction de leur zone: elles sont plus élevées pour les enfants vivant dans des zones à plus forte concentration de travail des enfants.

Dans ce type de conception, une option couramment utilisée consiste à imposer une limite au nombre maximum d'unités finales qui seront sélectionnées dans n'importe quelle zone. Pour une situation, un plan et une taille d'échantillon cible donnés, la limite supérieure détermine le nombre de zones à sélectionner dans l'échantillon – et vice versa, le nombre de zones détermine la limite appropriée pour assurer la taille d'échantillon souhaitée.

3.9. Comment traiter les très grandes unités

Dans le contexte de l'échantillonnage avec PPT, «très grand» s'applique à une unité dont la mesure de taille dans l'équation [1] excède l'intervalle d'échantillonnage, c'est-à-dire, $P_i > I$, si bien que la probabilité impliquée de sélection $f1_i$ excède 1,0, ce qui est impossible. Plusieurs méthodes permettent de traiter ce problème:

1. Les grandes unités de la base de sondage peuvent être segmentées (divisées en unités plus petites) de telle sorte qu'aucun segment n'excède la taille de I. Les segments constituent ensuite les unités d'échantillonnage appropriées.

2. La division de grandes unités peut n'être que conceptuelle. Une grande unité est considérée comme si elle était faite de t unités, selon la définition ci-dessous:

   $t = \mathrm{Int}(p_i/I)$ avec une probabilité $x = \mathrm{Int}(p_i/I) + 1 - (p_i/I)$, et

   $t = \mathrm{Int}(p_i/I) + 1$ avec une probabilité $(1 - x) = (p_i/I) - \mathrm{Int}(p_i/I)$

   L'unité est censée avoir été sélectionnée t fois, et t sous-échantillons séparés des unités finales sont sélectionnés à partir de celle-ci de la manière habituelle. Ce type de procédure est souvent utilisé le plus commodément avec un échantillonnage systématique. La variation aléatoire de la valeur de t rend la taille d'échantillon variable, mais les probabilités de sélection sont correctement préservées.

   On notera qu'avec l'échantillonnage systématique, il n'est pas nécessaire de déterminer réellement t, qui est une variable aléatoire, pour chaque grande zone. Quand les unités sont listées et que l'intervalle d'échantillonnage systématique est appliqué à leurs mesures de taille cumulatives, t est simplement le nombre de points de sélection qui se retrouvent sur la zone. Le point important est qu'une fois qu'une zone a été sélectionnée t fois de cette manière, ce n'est pas un seul mais plusieurs sous-échantillons t séparés qui doivent être sélectionnés à partir de celui-ci. On peut aussi sélectionner dans toute la zone un seul sous-échantillon, mais dont la taille sera t fois celle du sous-échantillon normal.

3. Traiter les grandes unités comme étant automatiquement sélectionnées. L'unité sectorielle est sortie de la base de sondage et supposée avoir été automatiquement sélectionnée (unité «autoreprésentée»); elle est ensuite sous-échantillonnée avec la probabilité de sélection globale requise. Cela donne $f1_i = 1; \quad f2_i = f.$

   Cette procédure revient à attribuer à chaque grande unité de taille p_i une nouvelle mesure de taille =I et à sélectionner l'échantillon de zones de manière systématique avec un intervalle I. Dans un plan autopondéré, la taille d'échantillon provenant de la zone est ensuite proportionnelle à sa taille actuelle, $b_i = f * p'_i$. .

4. Dans un plan à nombre fixe d'unités b=f*I sélectionnées par zone, on peut utiliser la procédure suivante. La grande unité est intégrée automatiquement dans l'échantillon et les b unités finales sont sélectionnées dans un second stade. Cela signifie

   f1i=1, f2i=b/p"i=f*(I/p"i)

   Les données doivent donc être pondérées positivement par (p'_i/I); il est clair qu'un tel plan n'est pas approprié si la taille de l'unité sectorielle est sensiblement supérieure à I.

La méthode (1) est en principe la meilleure, mais elle peut être coûteuse. Elle est nécessaire quand une proportion élevée des unités de la base est «très grande» au sens ci-dessus. La méthode (2) a parfois été utilisée par commodité, car elle évite de retirer les grandes unités de la base utilisée pour la sélection des échantillons; toutefois, elle accroît la variabilité de la taille d'échantillon finale. La méthode (4) n'est raisonnable que si aucune taille p_i n'excède largement la limite I.

La méthode (3) est la procédure normalement recommandée.

3.10. Comment traiter les très petites unités

Dans le contexte de l'échantillonnage avec PPT, «très petite» s'applique à une unité dont la mesure de taille dans l'équation [2] est plus petite que le tirage requis, c'est-à-dire que $p_i < b$, si bien que la probabilité impliquée de sous-échantillonnage $f2_i$ excède 1,0, ce qui est impossible. Plusieurs méthodes permettent, ici encore, de traiter ce problème:

1. Les petites unités de la base de sondage peuvent être regroupées (fusionnées pour constituer de plus grandes zones), de telle manière qu'aucun groupe ne soit plus petit que le tirage d'échantillon b requis. Chaque groupe constitue ensuite une UPE appropriée.

   Pour maintenir une contiguïté géographique, les petites unités sont souvent fusionnées avec des unités voisines, qu'elles soient petites ou grandes. Cela exige une connaissance de la localisation physique et des limites des unités dans la base, qui n'est pas toujours disponible dans les listages d'unités de zone non accompagnés de cartes. Toutefois, il est fréquent que les listes disponibles soient ordonnées géographiquement, ce qui peut aider à identifier approximativement les unités contiguës.

   Une solution de substitution consiste à composer une strate séparée de petites unités, puis à sélectionner simplement un échantillon à probabilité égale dans celle-ci. Toutefois, cela ne permet pas de contrôler la répartition géographique des petites unités de l'échantillon.

2. Il est souvent plus commode de sélectionner d'abord un échantillon des unités de zone originales de la manière habituelle, et de regrouper seulement ensuite les unités sélectionnées pour obtenir la taille minimale requise avant le stade suivant de l'échantillonnage.

   Cela est admissible quand les règles utilisées pour la formation de groupes sont objectives, en ce sens que le mode de groupement des unités est indépendant de la sélection de telle ou telle unité dans le premier échantillon. Ces règles sont faciles à établir.

3. Attribuer aux petites unités une certaine mesure de taille minimale. Une autre méthode consiste à attribuer simplement à toutes les unités une mesure de taille minimale =b. Cela signifie que l'on donne à toutes les unités de la base une taille minimale b au stade de la sélection des zones, quelle que soit leur taille réelle. Avec des unités pour lesquelles la taille originale était égale à ou plus petite que b, toutes les unités finales sont intégrées dans l'échantillon au dernier stade («échantillonnage recensement» ou «échantillonnage en grappe compacte»). Les probabilités de sélection finale initiales demeurent inchangées, mais, pour une taille d'échantillon donnée, le nombre des unités de zone de l'échantillon est accru. Cela peut ne pas être pratique et coûteux s'il y a trop de petites unités dans la base.

On notera que cette procédure signifie simplement que les unités classifiées comme très petites sont sélectionnées avec une probabilité constante plutôt qu'avec PPT. Elle est semblable à la démarche consistant à former une strate séparée de ces unités et à les sélectionner à un taux constant approprié.

4. Une des solutions au problème ci-dessus consiste à maintenir sans changement les mesures de taille des zones. Cela n'augmentera pas le nombre des unités de zone sélectionnées mais demandera une pondération positive des échantillons finals dans les unités avec $p_i<b$ par le facteur (b/p_i).

5. Exclure les plus petits des petites zones, avec une compensation appropriée. La méthode suivante s'est révélée utile dans un certain nombre d'enquêtes sur le travail des enfants. En présence d'un grand nombre de petites unités, une solution plus pratique comporte une combinaison de (3) et (4). Même si cette méthode est approximative, elle évite l'inclusion d'un trop grand nombre de très petites unités dans l'échantillon. Cela est particulièrement utile quand la population inclut de nombreuses unités, chacune d'elles avec seulement quelques unités finales intéressantes (voire même aucune) – ce qui peut fort bien être le cas dans les enquêtes sur le travail des enfants. Cette procédure est esquissée ci-dessous; on en trouvera une illustration chiffrée simulée dans la section suivante.

   i) Supposons que $\sum p_i$ est la mesure de la taille totale de l'ensemble de «très petites unités» à échantillonner, avec les paramètres de plan I comme intervalle d'échantillonnage à appliquer pour une sélection systématique des zones, f le taux d'échantillonnage global et b=f*I la taille d'échantillon attendue par zone de sondage. Selon ce plan, le nombre attendu d' unités finales qui devraient être sélectionnées dans cette strate est évidemment $f\cdot\sum p_i$.

   ii) Cet ensemble de petites unités est ordonné par taille d'unité et divisé en deux parties. La première partie consiste en $A = \Sigma p_i / b$ plus grandes unités de cet ensemble de petites unités. Ces unités sont sélectionnées comme dans la procédure (3) ci-dessus. La mesure de taille de chaque unité est portée à b, si bien que la sélection avec l'intervalle I donne un échantillon de $a = A.b/I = \Sigma p_i / I$ unités de zone. Toutes les unités finales de chaque zone sélectionnée sont maintenues dans l'échantillon.

   iii) La seconde partie est constituée par les plus petites unités de l'ensemble des petites unités. Pour des raisons de coût et de commodité, ces unités sont entièrement exclues de l'échantillon, même s'il ne s'agit pas là à proprement parler d'une procédure d'échantillonnage probabiliste.

   iv) Pour compenser cette exclusion, nous augmentons le poids accordé à chaque unité sélectionnée dans la première partie: toutes les unités finales d'une zone sélectionnée de taille p_i sont pondérées positivement par le facteur (b/p_i), tout comme dans la procédure (4) ci-dessus. Cela a simplement pour effet de porter à b=f*I le tirage pondéré de chaque zone sélectionnée, ce qui donne un équivalent de $a.b = f.\Sigma p_i$ pour les zones sélectionnées.

v) En (iv) ci-dessus, on trouve exactement le nombre attendu correct indiqué en (i)[14]. En fait, c'est pour réaliser cet équilibre que le nombre A de zones incluses dans la première partie a été déterminé comme en (ii) ci-dessus, $A = \Sigma p_i / b$.

Le plan à tirage fixe

Les méthodes (1) et (2) peuvent être appliquées comme ci-dessus. D'autres options sont évidemment impossibles pour les petites zones, $p_i < b$, puisque ces zones ne peuvent pas autoriser le tirage prédéterminé *b*, sauf dans les cas où la taille réelle a augmenté et égale ou excède b.

3.11. Une illustration chiffrée

3.11.1. Simulation d'une population d'unités de zone

Pour illustrer certains des points évoqués ci-dessus, nous incluons ici un exemple chiffré. Un petit ensemble de données a été produit statistiquement pour fournir un exemple raisonnablement réaliste d'un ensemble de zones avec les mesures de taille associées. La population initiale est constituée de quelques P=50 000 unités finales (ménages), dans près de 800 zones d'une taille moyenne d'environ 65 ménages. Les zones variaient considérablement en taille, de la plus petite constituée de moins de 10 unités à la plus grande constituée de plus de 1 000 unités. La répartition est assez typique des enquêtes réelles, fortement asymétriques vers la gauche avec beaucoup de petites zones et quelques très grandes zones.

Un échantillon systématique avec PPT de a=100 zones a été sélectionné, l'intervalle d'échantillonnage pour la sélection des zones étant I=(P/a)=500. Les données présentées dans le tableau 3.3 ne concernent que les 100 zones sélectionnées, et non la population tout entière de près de 800 zones.

Les colonnes du tableau 3.3 se détaillent comme suit. Les 100 unités sélectionnées dans l'échantillon avec probabilité proportionnelle aux mesures de taille p_i sont listées par ordre de taille. Les unités ainsi listées sont numérotées séquentiellement de 1 à 100 dans la col. [2], et les mesures de taille correspondantes figurent dans la col. [3].

Une limite minimale de 20 et une limite maximale de 500 sont imposées aux mesures utilisées pour les sélections d'échantillons. La limite inférieure correspond au nombre cible, b=20, des ménages à sélectionner par zone, et la limite supérieure à l'intervalle de sélection, I=500, utilisé pour l'échantillonnage systématique. Ces mesures modifiées pour les unités sélectionnées figurent dans la col. [4]. L'équation de sélection avec PPT utilisée pour la sélection de l'échantillon présenté est

$$f0_i = p_i / 500$$

[14] Le nombre réel des unités qui apparaissent dans l'échantillon peut varier selon les différences existent entre les mesures de taille p_i de la base et les tailles réelles des unités constatées au moment de la sélection des unités finales. Mais la procédure de sélection est spécifiée entièrement en fonction des mesures de taille de la base.

sous-échantillon de zones pour les enquêtes ETE/EET et seront examinés aux chapitres 4 et 5.

La col. [1] est basée sur la liste des zones de la population tout entière ordonnée par valeur de p_i. Elle indique le nombre cumulatif de zones de la population aux points où une zone a été sélectionnée. La col. [2] peut être considérée elle aussi comme une cumulation, mais limitée aux zones de l'échantillon. Ainsi, par exemple, les plus petits 30 % des zones de la population sont représentées par 10 % de l'échantillon. De même, les plus petits 53 % des zones de la population sont représentés par 20 % des plus petites zones de l'échantillon. Cela résulte de la méthode de sélection avec PPT.

Tableau 3.3: Illustration: échantillonnage des unités de zone avec PPT

«Premier échantillon»: par example pour ETE ou EET.
Paramétres:
a= 100 [nombre de zones dans l'échantillon]
b= 20 [nombre d'unités sélectionnées espéré/zone d'échantillonnage]
I= 500 [intervalle pour la sélection systématique des zones]
n= 2,000 [=a*b, taille d'échantillon attendue]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Fréquence cumulative du nombre de zones		Mesure de la taille de la zone		Probabilité de sélection d'une zone	Numéro de l'enfant travailleur dans la zone	
dans la population	échantillon	pi	pi modifié	f1 (%)	ci	ci/pi
3,0	1	8	20	4,00	9	0,44
6,0	2	11	20	4,00	5	0,26
9,0	3	12	20	4,00	9	0,44
12,1	4	16	20	4,00	0	0,00
15,1	5	17	20	4,00	3	0,14
18,1	6	18	20	4,00	1	0,03
21,1	7	18	20	4,00	12	0,62
24,1	8	20	20	4,00	5	0,24
27,1	9	21	21	4,21	0	0,00
30,0	10	22	22	4,37	18	0,81
32,8	11	24	24	4,72	0	0,01
35,3	12	25	25	5,07	11	0,45
37,7	13	28	28	5,56	1	0,03
39,9	14	28	28	5,57	0	0,02
42,0	15	30	30	6,02	6	0,21
44,1	16	33	33	6,55	5	0,15
45,9	17	34	34	6,75	0	0,01
47,7	18	35	35	6,97	11	0,31
49,4	19	36	36	7,16	4	0,11
51,1	20	38	38	7,52	17	0,45
52,7	21	39	39	7,81	8	0,21
54,2	22	40	40	7,94	1	0,01
55,8	23	40	40	8,04	0	0,00
57,3	24	43	43	8,57	20	0,46
58,7	25	43	43	8,63	27	0,62

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Fréquence cumulative du nombre de zones		Mesure de la taille de la zone		Probabilité de sélection d'une zone	Numéro de l'enfant travailleur dans la zone	
dans la population	échantillon	pi	pi modifié	f1 (%)	ci	ci/pi
60,1	26	44	44	8,80	23	0,51
61,4	27	46	46	9,23	18	0,39
62,8	28	48	48	9,58	23	0,48
64,0	29	50	50	10,03	7	0,15
65,2	30	52	52	10,39	1	0,01
66,4	31	52	52	10,40	0	0,01
67,5	32	53	53	10,58	11	0,20
68,7	33	57	57	11,33	9	0,16
69,7	34	59	59	11,85	0	0,00
70,8	35	60	60	11,97	53	0,89
71,8	36	64	64	12,71	11	0,17
72,7	37	65	65	13,02	12	0,18
73,6	38	67	67	13,35	11	0,17
74,5	39	68	68	13,66	2	0,03
75,4	40	69	69	13,81	3	0,04
76,3	41	69	69	13,88	1	0,02
77,2	42	70	70	13,91	61	0,87
78,0	43	72	72	14,42	3	0,04
78,9	44	72	72	14,48	35	0,48
79,7	45	73	73	14,65	51	0,69
80,5	46	74	74	14,81	20	0,27
81,3	47	79	79	15,72	41	0,52
82,1	48	79	79	15,82	7	0,09
82,9	49	83	83	16,60	74	0,90
83,6	50	92	92	18,37	1	0,01
84,3	51	92	92	18,38	20	0,21
84,9	52	97	97	19,36	22	0,23
85,5	53	97	97	19,50	15	0,16
86,2	54	100	100	20,04	93	0,92
86,8	55	103	103	20,57	73	0,71
87,4	56	107	107	21,50	16	0,15
87,9	57	112	112	22,39	21	0,19
88,5	58	114	114	22,81	25	0,22
89,0	59	114	114	22,89	20	0,18
89,5	60	115	115	22,98	68	0,60
90,0	61	124	124	24,76	27	0,21
90,5	62	128	128	25,53	0	0,00
91,0	63	138	138	27,62	3	0,03
91,4	64	156	156	31,19	16	0,11
91,8	65	157	157	31,48	117	0,74
92,2	66	159	159	31,89	1	0,01

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Fréquence cumulative du nombre de zones		Mesure de la taille de la zone		Probabilité de sélection d'une zone	Numéro de l'enfant travailleur dans la zone	
dans la population	échantillon	pi	pi modifié	f1 (%)	ci	ci/pi
92,6	67	163	163	32,55	51	0,31
92,9	68	171	171	34,28	22	0,13
93,3	69	175	175	34,94	11	0,06
93,6	70	181	181	36,29	40	0,22
94,0	71	186	186	37,10	23	0,12
94,3	72	189	189	37,87	20	0,11
94,6	73	196	196	39,29	190	0,97
94,9	74	200	200	40,03	1	0,00
95,2	75	204	204	40,73	8	0,04
95,5	76	213	213	42,66	55	0,26
95,8	77	213	213	42,67	1	0,01
96,1	78	229	229	45,79	107	0,47
96,4	79	229	229	45,86	42	0,18
96,6	80	237	237	47,36	9	0,04
96,9	81	238	238	47,53	2	0,01
97,1	82	240	240	47,96	7	0,03
97,4	83	246	246	49,30	1	0,00
97,6	84	299	299	59,79	268	0,90
97,8	85	304	304	60,76	36	0,12
98,0	86	331	331	66,29	220	0,66
98,2	87	345	345	68,99	1	0,00
98,4	88	358	358	71,68	26	0,07
98,5	89	368	368	73,67	162	0,44
98,7	90	383	383	76,58	2	0,01
98,9	91	424	424	84,73	0	0,00
99,0	92	425	425	85,06	382	0,90
99,2	93	527	500	100,00	292	0,58
99,3	94	566	500	100,00	92	0,18
99,4	95	659	500	100,00	0	0,00
99,5	96	681	500	100,00	60	0,12
99,6	97	719	500	100,00	383	0,77
99,8	98	898	500	100,00	0	0,00
99,9	99	908	500	100,00	9	0,02
100,0	100	1076	500	100,00	39	0,08

Dans l'échantillonnage avec PPT, la taille moyenne de la zone dans l'échantillon peut être beaucoup plus importante que la taille moyenne dans la population – dans notre exemple elle est d'environ 170 contre 65. Cette différence dépend de la variabilité des tailles des unités dans la population. En fait, le ratio de taille moyenne des zones entre l'échantillon et la population est donné par

$$\frac{\overline{B}_{\text{échantillon}}}{\overline{B}_{\text{population}}} = 1 + cv^2\left(B_i\right)$$

dans laquelle cv (B_i) est le coefficient de variation des tailles de chaque zone dans la population[14]. Dans notre illustration, le ratio d'échantillon à la taille moyenne des zones de population correspond à un cv de 1,3, qui indique un degré élevé de variabilité dans la taille des zones.

L'implication pratique de ce point est que l'échantillonnage avec PPT augmente le travail nécessaire au listage, à la sélection et au dénombrement dans l'enquête, par comparaison avec ce que l'on peut attendre de la taille moyenne des unités de zone de la population. Cette augmentation dépend de la variabilité des tailles unitaires, le système d'échantillonnage avec PPT introduisant dans l'échantillon un nombre disproportionné d'unités plus importantes, qui est évidemment compensé par une proportion correspondante plus faible d'unités finales sélectionnées dans l'échantillon, dans chaque zone plus importante, au stade suivant de l'échantillonnage.

Supposons que ce soit là le premier degré de la sélection d'un échantillon autopondéré à deux degrés avec une probabilité globale $f=f1 \cdot f2=4$ %. Pour le second degré, l'équation de sélection est $f2_i=20/p_i$, ce qui donne $f=p'_i/500 \bullet 20/p_i=4/100$.

Pour simplifier, nous supposerons généralement, dans les illustrations de ce chapitre et des suivants, que les tailles réelles des zones (p'_i) au moment du dénombrement sont les mêmes que les mesures de taille initiales utilisées dans la sélection des zones avec PPT. En réalité, ces quantités seront différentes, cela dépendant essentiellement de l'ancienneté de la base de sondage.

La taille d'échantillon des unités finales d'une grappe est $p'_i \bullet f2_i=20(p'_i/p_i)$, qui est une constante (=20) dans l'hypothèse ci-dessus.

3.11.2. Traiter des unités très petites et très grandes

Pour illustrer notre propos, nous avons supposé que le tirage cible par zone est b=20. Cela détermine la mesure de taille minimale qui peut être attribuée à n'importe quelle unité. (Il y a 7 unités sur 100 dans l'échantillon, numérotées de 1 à 7, dont les mesures de taille se situent au-dessous de ce niveau.) Les unités dont la taille va jusqu'à cette limite ont été sélectionnées avec une probabilité $f1_i=(b/l)=(20/500)=4$ %, et toutes les unités finales de chaque zone sélectionnée sont incluses dans l'échantillon ($f2_i=1,0$). Cela donne le taux d'échantillonnage global requis f_i aussi =4 %. Il s'agit là d'un taux

[14] On notera que cette formule concerne des «valeurs attendues», c'est-à-dire qu'elle vaut pour des quantités représentant une moyenne de tous les échantillons possibles d'un plan donné. Pour tout échantillon donné, les valeurs sont bien entendu sujettes à la variabilité de l'échantillonnage.

d'échantillonnage global constant, qui s'applique à toutes les unités finales de la population.

L'intervalle d'échantillonnage I=500 détermine la mesure de taille maximale qui peut être attribuée à n'importe quelle unité. (Il y a, dans l'exemple, 8 unités d'échantillon, numérotées de 93 à 100, dont la mesure de taille excède I.) Les unités dont la taille est égale ou supérieure à cette limite sont automatiquement incluses dans l'échantillon ($f1_i$=1,0), et elles sont échantillonnées au taux $f2_i$=(b/I)=(20/500)=4 % au dernier stade. Cette sélection donne aussi le taux d'échantillonnage global f_i =4 %. Le nombre des unités finales incluses dans l'échantillon à partir d'une telle zone excédera la taille d'échantillon cible b=20 dans la mesure où la mesure de taille initiale de la zone b_i excède l'intervalle d'échantillonnage I (500). Par exemple, la zone la plus grande (taille=1 076) donnera 20*1 076/500=43 unités finales à l'échantillon.

3.12. Les tirages non probabilistes

Dans le contexte des statistiques officielles, presque toutes les enquêtes doivent être basées sur la probabilité et sur des échantillons mesurables. Néanmoins, on rencontre couramment des formes diverses d'écarts par rapport aux échantillons probabilistes dans d'autres contextes, comme dans les études de marché et d'opinion, et cela avant tout pour des raisons de coût et de commodité. Il peut en aller de même pour des enquêtes particulièrement intensives et/ou restreintes sur le travail des enfants, comme les enquêtes sur des populations particulières de travailleurs enfants dans des zones ou des types d'activité spécifiques.

L'objectif de cette section est de décrire quelques-unes des méthodes d'échantillonnage non probabiliste que l'on rencontre communément, en notant leurs limites mais aussi leurs usages potentiels.

Il importe de souligner d'emblée que, selon la bonne pratique, on devrait s'efforcer d'obtenir des échantillons probabilistes. Tout le développement ci-dessous devrait être interprété dans ce contexte.

Limites et usages possibles des méthodes d'échantillonnage non probabilistes

Pour l'échantillonnage probabiliste, c'est la randomisation, et non les hypothèses sur la structure de la population, qui caractérise le processus de sélection. Par contre, l'échantillonnage non probabiliste ne comporte pas de sélection *aléatoire*. Dans l'échantillonnage non probabiliste, il est impossible d'estimer la probabilité d'inclusion d'une unité particulière dans l'échantillon, ou d'assurer à chaque unité une chance d'être incluse, ce qui rend impossible l'estimation de la variabilité ou de la distorsion de l'échantillonnage dans les résultats d'une enquête. Les échantillons non probabilistes ne peuvent pas dépendre des principes de la théorie des probabilités: il n'y a pas de «distribution d'échantillonnage» à utiliser quand on tire des résultats d'un échantillon. Au lieu de cela, nous devons utiliser quelques modèles de probabilité supposée. Il peut s'agir d'un modèle des caractéristiques sous-jacentes de la population, ou encore du processus de sélection.

Avec de tels échantillons, nous pouvons représenter bien ou mal la population, et il est difficile d'évaluer le résultat. La fiabilité ne peut pas être mesurée dans un échantillonnage non probabiliste; la seule manière de traiter la qualité des données est de comparer les résultats d'enquête avec l'information sur la population disponible à partir d'autres sources.

Cela signifie-t-il que les échantillons non probabilistes ne peuvent fournir aucune information utile? Non, pas nécessairement. En dépit de ces inconvénients, les méthodes d'échantillonnage non probabiliste peuvent être utiles. Il peut se trouver des circonstances dans lesquelles il n'est pas faisable, raisonnable d'une manière pratique et théorique de procéder à un échantillonnage aléatoire. De plus, les méthodes non probabilistes sont souvent rapides, peu coûteuses et commodes. En voici les avantages possibles:

- Elles sont au moins utilisables quand l'échantillonnage probabiliste n'est pas possible – par exemple quand on ne dispose pas d'une base de sondage
- Elles sont normalement moins coûteuses et plus rapides. Ces avantages peuvent être particulièrement utiles quand une population est importante ou largement dispersée.
- Elles sont souvent utiles dans les études exploratoires, comme pour la formulation d'hypothèses ou la conception de questionnaires. Ainsi, dans les tests de questionnaires, nous pouvons davantage chercher à nous faire une idée de la série des réponses qu'à établir le pourcentage de la population qui apporte une réponse particulière.

Quand les méthodes non probabilistes peuvent être utiles

Il peut y avoir trois types de raisons d'utiliser une méthode non probabiliste pour la sélection d'un échantillon.

1. Un échantillon probabiliste est considéré comme souhaitable mais n'est pas réalisable.

Il s'agit là de situations dans lesquelles on n'a d'autre choix qu'entre l'utilisation d'un échantillon non probabiliste et le renoncement à l'enquête. Par exemple, il n'est pas possible d'utiliser une méthode d'échantillonnage strictement probabiliste quand une base de sondage appropriée n'est pas disponible pour certaines populations ou groupes particuliers de la population, comme une liste des enfants non scolarisés ou des ménages qui comportent de tels enfants. Ou encore quand le taux de non-réponses attendu est si élevé qu'un échantillon probabiliste ne peut pas être réalisé, même s'il a été sélectionné pour être tel. Ou quand le coût et le temps requis pour obtenir un échantillon probabiliste sont inaccessibles.

2. Un échantillon probabiliste est considéré comme inutile.

Cela s'applique à des situations dans lesquelles la population est relativement homogène (les unités ont des caractéristiques semblables) ou est bien mélangée (l'arrangement des unités dans la population est aléatoire), si bien que le mode de sélection exact de l'échantillon n'a pas beaucoup d'importance. Les méthodes d'échantillonnage dites «au hasard» décrites ci-dessous sont basées sur cette

hypothèse. Dans de telles situations, les méthodes de sélection non probabilistes peuvent suffire: elles sont choisies dans la mesure où elles sont plus rapides, moins onéreuses ou plus commodes.

3. Un échantillon non probabiliste est préférable

Cette situation doit être considérée comme rare, mais elle est assurément possible. Elle apparaît essentiellement dans des situations où certains types d'objectifs sont plus faciles à atteindre par une sélection des unités appuyée sur le jugement plutôt que par une procédure totalement aléatoire. On part de l'hypothèse selon laquelle la distorsion résultant d'un jugement d'expert est plus faible que la variabilité d'échantillons aléatoires de petite taille. Les méthodes d'échantillonnage dites «par choix raisonné» décrites ci-dessous sont basées sur cette hypothèse.

Les échantillons non probabilistes dans les agences de statistique officielles

La plupart des organisations statistiques officielles utilisent l'échantillonnage probabiliste pour leurs enquêtes, mais en général l'échantillonnage non probabiliste pour les tests de questionnaires et les études préliminaires au cours du stade de la mise au point d'une enquête. Dans un certain nombre de pays, les grandes enquêtes officielles par sondage sur les entreprises utilisent la sélection par choix raisonné à cause des problèmes sérieux rencontrés pour obtenir la coopération des enquêtés. Certaines parties de l'indice des prix à la consommation sont souvent basées sur l'échantillonnage non probabiliste (par exemple dans la sélection des points de vente), et cela essentiellement pour des raisons de coût. Autre exemple: de nombreuses enquêtes officielles sur les dépenses des touristes dépendent d'échantillons par la méthode des quotas (sur la base du pays de résidence, de la durée du voyage, du profil démographique, de l'aéroport ou de la gare, etc.).

Les écarts par rapport à l'échantillonnage probabiliste peuvent être plus acceptables dans des situations où l'information est utilisée essentiellement par un seul client ou par un ensemble de clients très restreint. Toutefois, dans le secteur public, il n'existe pas d'utilisateur unique bien défini et il est hautement souhaitable que la méthode d'échantillonnage utilisée soit largement acceptable. Les méthodes d'échantillonnage aléatoire assurent cette large acceptabilité. Les statisticiens officiels ont la responsabilité de générer des données qui pourront être utilisées par la société tout entière: ils ne doivent tolérer aucune distorsion contrôlable dans leurs produits et devraient effectuer les enquêtes par sondage en utilisant des méthodes d'échantillonnage probabiliste.

Les différentes méthodes d'échantillonnage non probabiliste en usage

On rencontre dans la pratique un large éventail de méthodes d'échantillonnage non probabiliste, et les étiquettes utilisées pour les identifier et les distinguer ne sont pas tout à fait normalisées. Nous chercherons ci-dessous à les catégoriser un peu plus systématiquement et à donner une brève description des plus courantes.

3.12.1. L'échantillonnage au hasard

L'échantillonnage au hasard consiste à *inclure des unités dans l'échantillon sans structure, sans règles et sans procédures déterminées*. Comme on l'a indiqué, l'hypothèse de base de cette approche est que la population est relativement homogène ou est bien mélangée, de sorte que le mode de sélection exact de l'échantillon importe peu.

L'échantillonnage de commodité

Des unités particulières sont incluses dans l'échantillon pour la seule raison qu'il est plus commode de procéder ainsi que d'en choisir d'autres.

Un tel échantillon n'est pas normalement représentatif de la population cible, car les unités de l'échantillon sont seulement sélectionnées si elles sont d'accès facile et commode.

Il y a bien des moments où une personne ordinaire utilise l'échantillonnage de commodité dans les activités normales de la vie quotidienne. Autre exemple: les journalistes de télévision recherchent souvent des «entretiens avec les gens de la rue» pour connaître leur point de vue sur une question.

L'avantage évident est que la méthode est facile à utiliser, mais il est contrebalancé par le grand danger de distorsion. Bien que les applications utiles de cette technique soient limitées, elle peut donner des résultats exacts quand la population est homogène ou bien mélangée.

L'échantillonnage volontaire

Comme l'implique cette expression, ce type d'échantillonnage intervient quand des personnes proposent volontairement leurs services pour une étude. Dans les expériences psychologiques ou les essais pharmaceutiques, par exemple, il serait difficile, et même contraire à l'éthique, de recruter des participants provenant du grand public. Dans ces cas-là, l'échantillon est prélevé sur un groupe de volontaires. Dans la pratique clinique, nous pouvons utiliser les clients qui sont à notre disposition comme échantillon. Les médias radiotélévisés utilisent souvent les «appels des auditeurs» pour interroger un public sur ses opinions. Dans de nombreux contextes de recherche, nous échantillonnons simplement en demandant des volontaires. Parfois, le chercheur propose un paiement pour attirer les répondants.

Le problème posé par tous ces types d'échantillons est évident: rien ne nous prouve que les répondants soient représentatifs de la population cible, et il est clair dans bien des cas qu'ils ne le sont pas.

Échantillonner des participants volontaires, par opposition à la population générale, peut introduire de fortes distorsions. Par exemple, dans les sondages d'opinion, il est fréquent que seuls tendent à répondre, dans un sens ou dans l'autre, les personnes qui se préoccupent fortement du sujet évoqué. On peut supposer que, de manière générale, les personnes qui participent à ces enquêtes ont des opinions différentes de celles des non-participants.

L'échantillonnage en boule de neige

Dans l'échantillonnage en boule de neige, on commence par identifier une personne qui répond aux critères d'inclusion dans l'étude. On lui demande ensuite d'en recommander d'autres dont elle sait qu'ils répondent aux mêmes critères. L'idée de base est d'élargir l'échantillon sélectionné à l'origine en permettant à ces unités d'introduire dans l'échantillon d'autres unités ayant un lien avec elles. Les procédures peuvent aussi être poursuivies avec l'objectif de construire finalement une base d'échantillonnage de la population étudiée. L'application la plus courante consiste toutefois à répéter ce processus jusqu'à ce qu'un nombre suffisant de cas ait été généré dans la population aux fins de l'enquête. *L'accent est souvent mis sur l'obtention d'un nombre de cas suffisant, et moins sur le fait que ces unités puissent constituer un échantillon représentatif.*

L'échantillonnage en boule de neige est particulièrement utile quand on essaie d'atteindre des populations qui sont inaccessibles ou difficiles à découvrir. Par exemple, si l'on étudie les sans-abri, on a peu de chances de trouver des listes exactes de personnes sans abri dans une zone géographique spécifique. Toutefois, si l'on se rend dans cette zone et qu'on identifie quelques personnes sans abri, on constatera peut-être qu'elles savent très bien qui sont les autres personnes sans abri dans leur voisinage, et comment on peut les trouver.

Une condition nécessaire pour l'application de cette procédure est que les membres de la population soient en relation ou se connaissent. De par la nature de ce type d'échantillonnage, les membres de la population rare qui ont de nombreux contacts avec d'autres membres de cette population tendent à être sur-représentés, et ceux qui sont isolés des autres tendent à être sous-représentés dans l'enquête. La procédure a donc peu de chances de produire un échantillon représentatif, bien qu'elle puisse parvenir à fournir des informations utiles et constituer, dans certaines situations, la seule option disponible pour étudier la population insaisissable.

3.12.2. L'échantillonnage sur jugement

L'expression *échantillonnage sur jugement* est utilisée par référence à la sélection d'unités sur la base d'un jugement selon lequel les unités sélectionnées sont d'une certaine manière «représentatives» de la population concernée (ou de certains de ses aspects spécifiques).

Cette méthode est souvent utilisée dans les études exploratoires comme les prétests de questionnaires, les groupes thématiques, dans des contextes de laboratoire où le choix de sujets expérimentaux reflète les convictions préexistantes de l'enquêteur sur la population. L'hypothèse sous-jacente est que l'enquêteur sélectionnera des unités qui seront caractéristiques de la population pour les besoins de l'enquête. La question cruciale est ici celle de l'objectivité: dans quelle mesure peut-on se fier au jugement pour parvenir à un échantillon typique ?

L'échantillonnage par choix raisonné

Cette procédure peut être utilisée quand le nombre d'unités à sélectionner est si faible que la variabilité avec sélection aléatoire sera excessivement importante, et

potentiellement plus dommageable que la distorsion inhérente à la sélection sur jugement. Le jugement porte sur la décision d'inclure ou non des unités particulières dans l'étude, plutôt que sur un mécanisme de sélection d'unités à inclure dans l'échantillon.

On peut en donner comme exemples les études basées sur un très petit nombre de zones ou de sites. L'exemple classique cité dans les manuels d'échantillonnage est le cas où l'on souhaite obtenir des informations sur la population urbaine, mais où l'on n'a les moyens d'échantillonner que dans une seule grande ville. Dans ce cas, les manuels indiquent clairement qu'il serait préférable de ne pas utiliser l'échantillonnage probabiliste pour choisir la ville. Ils suggèrent de faire preuve de jugement pour sélectionner une ville «représentative». Toutefois, si les zones incluses peuvent être déterminées sur la base du jugement, la sélection des unités finales (ménages, personnes, enfants, etc.) au sein de chaque zone incluse peut être aléatoire.

Cette procédure suppose que le phénomène d'intérêt au sein de la population générale est représenté, et peut être capté, dans un nombre restreint d'unités soigneusement choisies. On suppose que la variabilité essentielle réside dans (et non pas entre) les unités dont un nombre très limité seulement ont été incluses dans l'étude.

La localisation restreinte de l'échantillonnage

Elle s'applique quand une partie notable de la sous-population concernée est considérée comme confinée dans un nombre limité de localisations connues. L'échantillon est souvent limité à ces seules localisations sur la base de ce type de jugement ou d'appréciation. Cela peut réduire considérablement le coût de l'enquête, mais une distorsion est introduite, dans la mesure où la partie non couverte de la population établie en dehors de la zone de concentration a des caractéristiques différentes de celles de la partie couverte dans les zones de concentration.

Malgré la nécessité de réduire le nombre des localisations à inclure dans l'échantillon, il est souhaitable de maintenir la nature probabiliste de l'échantillon dans toute la mesure du possible, par exemple en utilisant l'échantillonnage aléatoire d'unités individuelles au sein de chaque localisation intégrée dans l'échantillon restreint.

L'échantillonnage par instance modale

Le mode est la valeur dont l'occurrence est la plus fréquente dans une répartition. Quand nous établissons un échantillon par instance modale, nous essayons d'inclure le type de cas le plus fréquent ou typique dans la population. De nombreux sondages d'opinion, par exemple, affirment interroger des électeurs «typiques». Cette approche de l'échantillonnage pose un certain nombre de problèmes. Par-dessus tout, comment savoir ce qu'est le cas «typique» ou «modal» ? Il est clair que l'échantillonnage par instance modale n'est raisonnable que dans des contextes d'échantillonnage informels.

L'échantillonnage par les extrêmes

Il s'agit d'un échantillonnage sur jugement dans lequel on recherche les cas extrêmes, par exemple pour comprendre ou mettre en lumière des facteurs sous-jacents, des causes, des conséquences, etc. Cette sorte d'approche peut être utile pour enquêter, par exemple, sur les pires formes de travail des enfants.

3.12.3. L'échantillonnage par quotas et autres «structures»

Il s'agit de formes d'échantillonnage non probabiliste dans lesquelles la sélection d'unités dans la population s'effectue par la médiation d'une structure ou d'un ensemble de contraintes. Toutefois, au sein de cette structure, les sélections ne sont pas aléatoires. L'étroitesse de la structure détermine le degré de proximité de l'échantillon obtenu par rapport à un échantillon probabiliste. Un exemple de cette allocation de quotas d'échantillon à différentes parties de la population est décrit ci-dessous.

L'échantillonnage par quotas

C'est l'une des formes les plus courantes d'échantillonnage non probabiliste.

L'échantillonnage par quotas implique une sélection assortie de contrôles garantissant que des nombres d'unités spécifiés (quotas) sont obtenus de chaque sous-groupe spécifié de la population (comme les ménages ou les personnes classifiées selon les caractéristiques appropriées), mais cela, pour l'essentiel, sans randomisation de la sélection des unités au sein des sous-groupes. L'idée maîtresse de l'échantillonnage par quotas est de produire un échantillon harmonisé avec la population cible sur certaines caractéristiques (ex.: âge ou sexe) en remplissant des quotas pour chacune de ces caractéristiques. L'hypothèse sous-jacente est que si l'échantillon s'harmonise avec la population sur ces caractéristiques, il peut aussi s'harmoniser avec la population cible sur les quantités que l'on cherche à mesurer. On notera que cette méthode exige de disposer de données exactes sur la population tout entière afin de fixer les quotas. Ainsi, si nous fixons des quotas par âge et par sexe, nous devons connaître la répartition de la population par âge et par sexe.

L'hypothèse qui dirige la procédure est que la variabilité essentielle existe entre les sous-groupes choisis plutôt qu'à l'intérieur de ceux-ci: dès lors que des groupes suffisamment petits et homogènes ont été définis et représentés de manière adéquate, il importe donc assez peu de savoir quelles unités individuelles particulières sont dénombrées au sein d'un groupe.

L'échantillonnage s'effectue jusqu'à sélection d'un nombre spécifique d'unités (quotas) pour diverses sous-populations. Comme il n'existe pas de règles qui déterminent la manière de fixer ces quotas, l'échantillonnage par quotas est en fait un moyen d'atteindre des objectifs de taille d'échantillon pour certaines sous-populations. Il peut être considéré comme préférable à d'autres formes d'échantillonnage non probabiliste (par ex. l'échantillonnage sur jugement) parce qu'il impose l'inclusion de membres de différentes sous-populations.

L'échantillonnage par quotas est souvent utilisé dans les études de marché à la place de l'échantillonnage aléatoire stratifié. Cette méthode est habituellement justifiée en termes de commodité, de rapidité et d'économie. Elle est plus facile à administrer et présente l'avantage de répondre aux proportions de la population. Elle évite en particulier les tâches de listage de la population tout entière, de sélection aléatoire de l'échantillon et de suivi des non-répondants. L'échantillonnage par quotas est une méthode efficace quand l'information doit être obtenue rapidement. Il peut être la seule méthode appropriée disponible quand il n'existe pas de liste convenable de la population à étudier.

Toutefois, cette méthode dissimule un biais potentiellement important. Comme pour toutes les autres méthodes d'échantillonnage non probabiliste, il est nécessaire, pour parvenir à des déductions concernant la population, de supposer que les personnes sélectionnées sont semblables à celles qui ne le sont pas. Des hypothèses aussi catégoriques sont rarement valables.

Bien que les enquêteurs soient soumis à la contrainte des quotas, ils utilisent souvent un élément de jugement dans le choix de l'échantillon. *Le degré de flexibilité dont disposent les enquêteurs varie d'une enquête à l'autre, et les règles et directives qui concernent cette flexibilité déterminent la mesure dans laquelle l'échantillon par quotas s'écarte de l'échantillonnage stratifié probabiliste.*

De même qu'il existe de nombreux modèles d'échantillons probabilistes, il existe plusieurs méthodes d'échantillonnage par quotas. Un échantillon par quotas peut être établi par degrés successifs. Il est courant, mais non pas nécessaire, que l'échantillonnage par quotas utilise des procédures de sélection aléatoires dans ses premiers stades, de manière très semblable au processus d'échantillonnage probabiliste. Par exemple, le premier stade d'un échantillonnage à degrés multiples pourra être la sélection aléatoire des zones géographiques. La différence réside dans la sélection des unités aux stades finals du processus.

Comparaison entre l'échantillonnage probabiliste et l'échantillonnage par quotas

Les principales différences entre l'échantillonnage probabiliste et l'échantillonnage par quotas sont les suivantes:

1. Si l'échantillonnage probabiliste est convenablement effectué, il n'y aura aucun des biais qui peuvent résulter de jugements subjectifs dans la sélection des échantillons. Toutefois, de tel biais sont possibles dans les échantillons par quotas. Ainsi, les enquêteurs peuvent choisir, consciemment ou non, des répondants non menaçants ou d'approche facile, ou bien ceux qu'il est aisé de contacter.

2. La méthode des quotas exige la formulation d'un modèle hypothétique adapté aux données. Par contre, une enquête probabiliste ne dépend en principe d'aucun modèle. La validité du modèle sous-jacent à l'échantillonnage par quotas peut être sujette à caution et difficile à vérifier.

3. Avec l'échantillonnage probabiliste, nous utilisons des procédures statistiques pour tirer des conclusions de l'échantillon et des erreurs d'échantillonnage. Dans un échantillon par quotas, nous ne pouvons pas obtenir des estimations de précision comparables.

4. Le plus souvent, la non-réponse dans un échantillon par quotas est traitée simplement par sélection d'un autre répondant approprié au quota. Généralement, la non-réponse peut être traitée plus efficacement dans un échantillon à base probabiliste.

5. De manière générale, le coût d'un échantillon par quotas est plus faible que celui d'un échantillon à base probabiliste de la même taille. Cependant, la question du coût est plus complexe et ne peut être considérée isolément de celle de la qualité des données. «Il n'existe aucun moyen de comparer le coût d'un échantillon

aléatoire avec celui d'un échantillon basé sur un jugement, car ces deux types d'échantillon sont utilisés à des fins différentes. Le coût ne signifie rien sans appréciation de la qualité, et il est impossible d'apprécier objectivement la qualité d'un échantillon basé sur un jugement comme celle d'un échantillon probabiliste.» (Edward Deming).

L'échantillonnage axé sur l'hétérogénéité

Nous procédons à un échantillonnage axé sur l'hétérogénéité quand l'objectif consiste à capter toute la complexité d'un phénomène (par exemple des opinions ou des points de vue), mais que l'on ne se préoccupe pas de représenter ces opinions ou les individus de manière proportionnelle. On parle aussi, dans ce cas, d'échantillonnage axé sur la diversité. Dans de nombreux processus de groupe (séances de réflexion et autres), par exemple, nous utilisons une forme d'échantillonnage axé sur l'hétérogénéité quand notre préoccupation majeure est d'obtenir un large éventail d'idées, et non d'identifier les idées «moyennes» ou «modales»; en effet, nous ne cherchons pas à échantillonner des personnes, mais des idées. Nous imaginons qu'il existe un univers de toutes les idées possibles en rapport avec un thème donné et que nous tenons à échantillonner cette population-là, et non celle des personnes qui ont ces idées. Il est clair que pour réunir toutes les idées, et notamment celles qui sont extrêmes ou inhabituelles, nous devons inclure un éventail de participants vaste et divers. En ce sens, l'échantillonnage axé sur l'hétérogénéité est presque l'opposé de l'échantillonnage par instance modale décrit ci-dessus.

Ce type d'échantillonnage peut revêtir la forme d'un échantillon par quotas non proportionnels un peu moins restrictif. Dans cette méthode, le nombre minimal des unités à échantillonner dans chaque catégorie est spécifié. On ne se préoccupe pas d'obtenir des nombres harmonisés aux proportions existant dans la population.

L'échantillonnage in situ

Cette technique est utilisée pour échantillonner des populations définies sur la base de l'activité (ou de la situation) dans laquelle elles sont engagées (ou dans laquelle elles se trouvent) sur certaines localisations spécifiques. On en trouve des exemples dans les enquêtes sur les passagers dans les aéroports ou les gares routières ou ferroviaires, les enquêtes sur les visiteurs dans les musées, les centres commerciaux, etc.

Généralement, ce type d'échantillonnage implique la sélection de localisations, de périodes d'observation, d'unités individuelles, et éventuellement aussi d'un sous-ensemble d'activités ou d'états à observer.

Une distinction importante porte sur le fait de savoir si l'enquête est conçue pour l'estimation d' «unités» ou d'«activités» en tant qu'éléments d'analyse. Un exemple peut en être donné par le nombre de visiteurs – par opposition au nombre de visites – dans un centre sanitaire. La probabilité de sélection d'une unité dépend du nombre des «activités» grâce auxquelles l'unité peut être dénombrée. Des procédures spéciales sont nécessaires quand un échantillon des premiers éléments (unités) est construit à partir des derniers (activités, etc.). Par contre, à des fins multiples, l'échantillonnage sur les localisations est essentiellement effectué en vue d'étudier des activités ou des

visites, etc., plutôt que des individualités, ce qui est plus simple en termes d'échantillonnage.

Les difficultés courantes dans ce type d'enquêtes concernent les conditions défavorables et les limites dans le temps qui peuvent s'opposer à la sélection d'échantillons et à la collecte de données pour le dénombrement de flux sur des localisations déterminées. Cela peut également occasionner des distorsions considérables dans la sélection et dans les mesures.

Chapitre 4

L'enquête sur le travail des enfants (ETE): l'échantillon intégré de taille réduite

Selon la définition du chapitre 2 (section 2.1.1), nous utilisons l'expression *enquête sur le travail des enfants* (ETE) par référence spécifique à une enquête ou à une composante d'enquête conçue avec l'objectif essentiel de mesurer l'ampleur du travail des enfants, éventuellement en incluant, le cas échéant, des informations sur la variation de l'ampleur en fonction des divisions géographiques/administratives, du type de lieu et de différentes caractéristiques des ménages et des individus.

La complexité de la collecte des données et des questionnaires requis à cet effet dépend de l'ampleur et de l'exhaustivité de l'information nécessaire pour identifier le travail des enfants selon la définition adoptée dans l'enquête en question.

Pour définir la structure d'une ETE, nous devons considérer ses liens avec les opérations qui la précèdent, si elles existent, et avec celles qui la suivent. Ce chapitre est consacré aux liens de l'ETE avec les opérations qui la précèdent. Trois dimensions sont à étudier: (1) l'opération précédente peut être une opération de listage des ménages, ou bien une enquête à grande échelle comme l'EMO; (2) l'ETE peut être combinée avec cette opération, ou bien être menée ultérieurement comme une opération séparée; et (3) l'ETE peut être conduite sur le même échantillon, ou sur un sous-échantillon de l'opération précédente. Les combinaisons de ces trois dimensions sont exposées ci-dessous. Bien entendu, certaines de ces combinaisons sont plus probables (plus significatives, plus pratiques que d'autres).

Liens d'une ETE avec l'opération qui la précède				
Intégration ou séparation des deux opérations				
Opération précédente	**Intégrée**		**Séparée**	
Listing des ménages	ETE comportant de brèves questions éliminatoires	même échantillon (complet)	ETE autonome	utilisation d'un échantillon de toute la liste improbable
		sous-échantillonnage improbable		sous-échantillonnage
EMO ou analogue	ETE modulaire ou «combinée»	même échantillon (complet) pour l'essentiel	ETE intégrée	parfois même échantillon (complet)
		parfois sous-échantillonnage		sous-échantillonnage pour l'essentiel

L'intégration de l'ETE avec une opération de listage des ménages exige qu'elle ne comporte que quelques brèves questions qui puissent être incorporées dans le formulaire de listage. Tous les ménages au sein de chaque échantillon doivent être listés, et également tout sous-échantillonnage est généralement impossible pour les questions d'ETE. Les données résultantes peuvent présenter l'avantage d'être basées sur un vaste échantillon, mais la mesure du travail des enfants sera probablement

approximative. Il est plus approprié de considérer ce type d'ETE comme une *opération de triage* (voir section 2.3).

L'enquête sur le travail des enfants peut être une enquête «indépendante», généralement basée sur une opération de listage des ménages, et sans lien avec une autre enquête telle qu'une EMO. Un sous-échantillonnage au sein des zones listées sera généralement nécessaire. Dans cette situation, les principes généraux d'échantillonnage valables pour différents types d'enquêtes sur la population qui ont été examinés au chapitre précédent s'appliquent également à l'ETE. Toutefois, une enquête indépendante sur le travail des enfants n'est pas toujours une option réalisable ni même souhaitable.

Une solution de substitution consiste à baser l'ETE, du moins en matière d'échantillonnage, sur une autre enquête auprès des ménages plus vaste, telle que l'EMO. Dans cette situation, il est courant d'incorporer l'ETE dans l'enquête de base (EMO) en tant que *module* (ou sur la forme d'une enquête «combinée» au sens de la description de la section 2.2.3). Généralement, cette disposition, qui comporte une intégration opérationnelle, implique que le module d'ETE soit appliqué sur le même échantillon de ménages que l'enquête de base. Il est toutefois possible de pratiquer le sous-échantillonnage au niveau sectoriel, autrement dit d'introduire l'ETE, en tant que module, seulement dans un sous-échantillon des zones de l'EMO. Une forme particulièrement commode de cette méthode consiste à inclure l'ETE seulement dans quelques phases d'une EMO continue.

Une autre solution de substitution est celle de l'*ETE intégrée,* au sens qui est exposé dans la section 2.2.3 – en tant qu'enquête dépendant de l'enquête de base pour son échantillon et éventuellement pour d'autres informations fournies, mais, pour le reste, est séparée opérationnellement de celle-ci. Ici, un sous-échantillonnage plus élaboré à partir de l'enquête de base est possible, que ce soit au niveau sectoriel ou à celui des ménages.

Pour des raisons qui seront explicitées dans la section 4.1.3, nous estimons que *la dernière option mentionnée, «ETE intégrée avec sous-échantillonnage», mérite davantage de considération qu'elle n'en a apparemment reçu dans les enquêtes passées sur le travail des enfants.*

Ce chapitre décrit les détails techniques des procédures d'établissement de l'échantillon d'une enquête sur le travail des enfants (ETE) sur la base de l'échantillon utilisé pour une enquête plus vaste sur la population générale, en particulier l'EMO. Les deux enquêtes tendent à présenter des conditions semblables concernant la structure de base de l'échantillon, mais elles sont censées différer par certains paramètres de plan et par la forme des liens existant entre elles. Ces différences ne doivent pas être oubliées quand on décide des procédures de sous-échantillonnage décrites ci-dessous.

4.1. La structure de l'ETE et ses liens

4.1.1. La structure commune

En mesurant l'incidence du travail des enfants, la population de base d'intérêt dans une ETE est la *population des enfants exposés au risque de travail des enfants*. Cette population de base est définie essentiellement en termes de limite d'âge, qui tendent à être bien réparties dans la population générale. Même dans des zones limitées, la taille de cette population est, dans la plupart des cas, approximativement proportionnelle à la population totale. Par conséquent, si l'indicateur de taille est, dans une EMO, constitué par la population générale (ou par la population en âge de travailler), il est donné, pour une ETE, par une population d'enfants exposés au risque du travail des enfants définie de manière appropriée. Il se trouve généralement que ces deux populations sont étroitement liées par leur taille, la différence moyenne entre l'une et l'autre étant essentiellement un facteur de mise à l'échelle.

Ces similitudes dans la structure de base des échantillons ont des implications importantes pour le choix d'une structure d'enquête appropriée pour l'ETE. Deux aspects essentiels en sont: la répartition de l'échantillon entre les domaines ou les strates de la population; et la sélection des zones de sondage.

La répartition de l'échantillon

Abstraction faite de toute répartition disproportionnée de l'échantillon visant à répondre à des exigences redditionnelles particulières (comme le suréchantillonnage des petits domaines répondant à des exigences de précision minimale), il est généralement approprié et courant, dans les enquêtes sur la main-d'œuvre, de répartir l'échantillon de manière proportionnelle, c'est-à-dire en fonction de la taille des domaines de population. En théorie, on peut «optimiser» la répartition en prenant en compte les différences de variance et de coûts entre les domaines. Mais les gains ainsi obtenus sont généralement trop faibles pour justifier un écart par rapport à la répartition proportionnelle plus simple et plus pratique. Cela s'applique particulièrement aux estimations de proportions (telles que le taux de chômage).

La situation est très semblable dans le cas d'une ETE sur l'ampleur du travail des enfants. C'est seulement en présence de différences très prononcées entre domaines qu'une répartition disproportionnée (optimale) peut se justifier. Pour l'estimation d'une proportion p, le taux d'échantillonnage théoriquement optimal est en proportion de $[p.(1-p)]^{1/2}$. Cette variation de la répartition est très peu sensible à la valeur de p, entre 0,3-0,5 pour p dans une fourchette de 0,1 à 0,9. Par conséquent, la répartition proportionnelle demeure généralement appropriée pour l'ETE, tout comme pour l'EMO.

La répartition requise entre les deux types d'enquête peut différer, mais avant tout à cause de différences dans les conditions redditionnelles pour les domaines sous-nationaux. (Voir section 4.1.2.)

La sélection des zones de sondage

La procédure de sélection des zones de sondage peut aussi être la même, dans la mesure où les tailles de la population de base appropriées pour les deux enquêtes sont presque proportionnelles l'une par rapport à l'autre. Ainsi, si la sélection avec probabilité proportionnelle à la taille p_i est appropriée à l'EMO, il est raisonnable de supposer qu'à des fins pratiques, les mêmes mesures de taille p_i utilisées dans l'EMO pour la sélection des zones avec PPT sont également appropriées à l'ETE pour la même raison. Par conséquent, une ETE portant sur l'incidence du travail des enfants dans la population des enfants demandera probablement une *structure d'échantillon semblable à celle d'une enquête sur la population générale* telle que l'EMO.

Mis à part la similitude de la structure de la population de base concernée, les deux types d'enquête tendent aussi à être semblables par le mode de collecte des données, ainsi que par des aspects fondamentaux (concepts, définitions, classifications, questions, période de référence, etc.).

4.1.2. Les paramètres de plan

En dépit de similitudes dans la structure et la distribution de la population à échantillonner, l'ETE et l'EMO diffèrent en règle générale par un certain nombre de conditions et de paramètres de plan. Dans de nombreux pays, l'EMO est une enquête bien établie, régulière ou même continue. L'ETE est plus probablement une enquête nouvelle ou récemment instituée, conduite au mieux de manière périodique et souvent sur une base seulement ponctuelle. Les ressources disponibles pour l'ETE sont probablement plus limitées, et souvent moins certaines. L'EMO est de plus en plus soumise à l'obligation de produire des estimations séparées pour les différentes régions et sous-populations du pays, et cela de manière plus fréquente (annuellement ou même trimestriellement), alors que, dans la plupart des cas, l'objectif essentiel de l'ETE doit encore être, en premier lieu, de produire de temps à autre des estimations au niveau national. En bref, *l'EMO peut être considérée comme une grande enquête extensive et régulière,* et par comparaison *l'ETE comme une enquête plus limitée, plus intensive et moins fréquente*. Par conséquent, en dépit de similitudes dans leur structure de base, les deux types d'enquête diffèrent par le choix des paramètres de plan, tels que la taille d'échantillon, le nombre de zones sélectionnées pour l'échantillon et la taille connexe d'échantillon par zone, la répartition de l'échantillon entre différents domaines en fonction des conditions redditionnelles différentes des deux enquêtes, les modalités de la stratification, et les unités finales pour lesquelles des données sont collectées dans ces enquêtes.

4.1.3. Les liens

Comme nous l'avons indiqué, il est en principe possible de conduire l'ETE sur un échantillon exactement identique à celui de l'EMO (à une enquête à grande échelle analogue). En fait, dans un certain nombre de pays, on a simplement ajouté des questions sur le travail des enfants, sous forme de module, à une EMO en cours. Ce type d'arrangement présente un certain nombre d'avantages en termes d'économie, de commodité et de durabilité. Les informations provenant des deux enquêtes (ou plutôt,

dans ce cas, des deux parties de la même enquête) peuvent être liées et analysées ensemble. Par contre, de sérieux inconvénients peuvent apparaître. Le contenu du module supplémentaire (l'ETE) doit être limité, et peut ainsi insuffisamment satisfaire aux exigences de données pour une évaluation exhaustive du travail des enfants. La charge des réponses dérivées de la combinaison des entretiens se trouve accrue, avec un effet négatif possible sur les taux de réponse et la qualité des réponses, notamment pour la composante principale (EMO). Le système entièrement intégré peut aussi être d'une rigidité excessive face aux exigences différentes pour les composantes EMO et ETE.

À l'autre extrémité, comme nous l'avons déjà mentionné, dans un certain nombre de cas, des enquêtes nationales sur le travail des enfants ont été menées en tant qu'activités séparées autonomes. Si l'enquête indépendante est susceptible de générer des données plus fiables sur le travail des enfants, il est cependant évident qu'il s'agit là d'une option coûteuse, souvent difficile à assurer ou à répéter.

Les enquêtes indépendantes sur le travail des enfants, exhaustives mais coûteuses, d'une part, et les modules complémentaires sur le travail des enfants, beaucoup plus accessibles et commodes, mais limités, d'autre part, constituent deux options possibles et très couramment utilisées.

Toutefois, dans de nombreuses situations, aucune de ces solutions extrêmes – intégration totale ou séparation totale – n'apporte la solution appropriée. Une solution plus pratique consiste souvent à chercher à lier l'ETE à une enquête à grande échelle sur la population générale, et préférablement à l'EMO. Ce lien peut revêtir des formes multiples et atteindre différents degrés.

En ce qui concerne les aspects d'échantillonnage de ces liens, une option commode et pratique consiste à tirer l'échantillon d'enquête sur le travail des enfants comme sous-échantillon d'une enquête plus vaste sur la population générale, la plus appropriée étant l'enquête sur la main-d'œuvre.

Bien entendu, il existe aussi un certain nombre d'autres possibilités. À un extrême, les deux enquêtes peuvent être basées sur des échantillons indépendants, mais même dans ce cas il est souhaitable et efficace d'extraire ces échantillons d'une base sectorielle commune, ou d'un «échantillon de référence» des zones, dont on peut tirer des échantillons en vue de différentes enquêtes. Cela permet le partage des coûts de préparation et d'entretien de la base sectorielle ou de l'échantillon de référence des zones.

Il est plus approprié, lorsque cela est possible, de baser les deux enquêtes sur le même échantillon de zones. En principe, toutes les zones de l'EMO peuvent être incluses dans l'échantillon d'ETE. Toutefois, le sous-échantillonnage des zones de l'EMO est souvent souhaitable et approprié, aussi bien du point de vue des objectifs d'enquête que de la disponibilité des ressources. Habituellement, les enquêtes sur la main-d'œuvre doivent fournir des estimations pour des domaines précis – différentes régions et autres divisions administratives, zones urbaines et rurales, différents sous-groupes démographiques et autres au sein de la population, etc. Par contre, on dispose souvent d'informations très limitées ou non existantes sur le travail des enfants, et initialement les résultats doivent être produits au niveau national, ou tout au plus seulement pour

quelques domaines majeurs. La taille d'échantillon pour l'ETE peut donc être beaucoup plus réduite. De même, une série échelonnée dans le temps de données basées sur des enquêtes régulières et fréquentes est généralement requise quand il s'agit de la main-d'œuvre générale, alors que la première des priorités en matière de travail des enfants est généralement de produire des informations fiables de nature plus structurelle. Les enquêtes sur le travail des enfants tendent donc à être des enquêtes uniques, tout au plus répétées occasionnellement, mais généralement sans liens temporels entre échantillons. En outre, comme nous l'avons déjà indiqué, les ressources disponibles pour ces enquêtes sont généralement plus limitées.

Au sein des zones de sondage communes, il existe différentes possibilités de relations entre les unités finales (par ex. les ménages) des deux échantillons, depuis les échantillons entièrement indépendants issus de listes des ménages dans les zones communes jusqu'aux échantillons d'ETE confinés aux enfants réellement identifiés au cours de l'entretien d'EMO. Nous examinerons d'abord les liens possibles au niveau des zones de sondage (Ceci est décrit dans la section 4.6).

4.2. La sélection des zones

Nous devrions maintenant examiner la procédure de sélection d'un nombre de grappes plus réduit pour l'échantillon d'ETE à partir d'un nombre plus important de zones de l'échantillon d'EMO, par exemple.

Les principes de ce sous-échantillonnage sont très simples, ainsi l'objectif de cette section et des suivantes est de fournir des précisions sur les procédures réelles, ainsi que quelques illustrations chiffrées en cas de besoin.

Dans cette section, nous commencerons par les situations les plus simples: pour un domaine donné, il faut sélectionner un nombre réduit de zones pour l'ETE à partir d'un échantillon donné de zones d'EMO. La section 4.3 traite de la situation où sont impliqués plusieurs domaines. Quelques complications supplémentaires interviennent quand nous avons affaire à des zones pour lesquelles des procédures spéciales ont été utilisées dans l'échantillon d'EMO existant à partir duquel doit être effectué le sous-échantillonnage. Il y a, par exemple, des zones qui ont été traitées, au sens évoqué dans les sections 3.9-10, comme étant «trop grandes» (avec des mesures de taille supérieures à l'intervalle d'échantillonnage I pour un échantillonnage systématique avec PPT), ou «trop petites» (avec des mesures de taille plus petites que le prélèvement d'échantillon cible b) au cours de leur sélection précédente. Ces problèmes particuliers sont examinés dans les sections suivantes. Nous nous préoccupons ici du sous-échantillonnage au niveau des zones; l'échantillonnage des ménages à l'intérieur des zones sera également examiné ultérieurement.

Il est un point pratique extrêmement important qu'il convient de noter quand on obtient un échantillon par sous-échantillonnage à partir d'un échantillon existant , à savoir toutes les précisions doivent être enregistrées concernant non seulement les procédures de sous-échantillonnage, mais aussi l'échantillon existant utilisé pour ce sous-échantillonnage. On trouve malheureusement des exemples d'enquêtes dans lesquelles les détails relatifs au plan de l'échantillon existant n'ont pas été correctement

documentés ou n'ont pas été conservés. L'élément d'information le plus crucial concerne les probabilités de sélection appliquées dans l'échantillon initial. Si ces précisions ne sont pas disponibles pour un échantillon donné, il est souhaitable de rechercher d'autres sources de sous-échantillonnage.

4.2.1. Sélection des zones avec probabilité constante dans un domaine

Supposons que dans un domaine de l'échantillon existant, a_1 zones aient été sélectionnées sur A_1 zones de la population, avec une probabilité constante $f_1=a_1/A_1$.

L'objectif est d'obtenir un sous-échantillon comptant moins de zones $a<a_1$, tout en conservant la même structure, c'est-à-dire une probabilité de sélection uniforme pour les zones $f=a/A_1$.

De manière évidente, le sous échantillonnage de zones requis est

$$g_1=a/a_1$$

Nous pouvons l'obtenir, par exemple, par un échantillonnage systématique à partir de l'échantillon existant avec probabilité constante en utilisant l'intervalle d'échantillonnage

$$k_1=1/g_1=a_1/a$$

Il est utile de considérer g_1 comme un ratio des taux d'échantillonnage ou des probabilités dans les deux échantillons, ce qui aide à généraliser exactement la même procédure à d'autres situations plus complexes, comme nous le verrons ci-dessous:

$$g_1=(a/A)/(a_1/A_1)=f/f_1$$

avec $A_1=A$ dans ce cas, bien entendu.

Le taux d'échantillonnage à appliquer est le *ratio probabilités ou taux d'échantillonnage nouveaux sur probabilités ou taux d'échantillonnage existants*.

De même, l'intervalle de sous-échantillonnage requis est $k_1=f_1/f$.

4.2.2. Sélection des zones avec probabilité proportionnelle à la taille dans un domaine

Supposons que l'EMO soit basée sur le procédé couramment utilisé de la PPT décrite au chapitre 3, et que cet échantillon contienne a_1 zones sélectionnées avec probabilité proportionnelle à une mesure de la taille de population (p_i). L'objectif est de choisir pour l'ETE un nombre de zones réduit, disons $a=a_1/k_1$, avec $k_1>1$, également avec probabilité proportionnelle à la même taille de la population p_i. En d'autres termes, on suppose donc ici que pour chaque zone i, la mesure de taille p_i utilisée dans la sélection initiale de l'EMO est la même que la mesure de taille à utiliser pour l'ETE. Il en va normalement ainsi quand les populations de base des deux enquêtes sont de proportions identiques ou proches et distribuées de manière semblable – par exemple tous les adultes pour l'EMO et les enfants de la même population pour l'ETE. La

situation dans laquelle on utilise dans les deux échantillons des types différents de taille sera examinée dans le chapitre 5.

Étant donné les mesures de taille communes, la procédure de base pour l'échantillonnage des zones de l'EMO vers l'ETE se déroule simplement comme suit:

- Choisissez un sous-échantillon de zones d'EMO avec une probabilité *constante* $g_1 = a/a_1$.
- La sélection avec probabilité constante (g_1) peut s'effectuer simplement en appliquant à l'EMO la procédure d'échantillonnage systématique à probabilité égale avec l'intervalle $k_1 = 1/g_1 = a_1/a$.

Le résultat obtenu est un échantillon avec PPT de *a* zones avec probabilité proportionnelle à la mesure de taille de la population p_i. Les équations de sélection pour les zones d'EMO et d'ETE sont les suivantes.

EMO: $f_1 = \left(\frac{a_1}{\Sigma p_i}\right).p_i = \frac{p_i}{I_1}$, par exemple, et

ETE: $f = \left(\frac{a}{\Sigma p_i}\right).p_i = \left(\frac{g_1 * a_1}{\Sigma p_i}\right).p_i = g_1 \cdot \left(\frac{p_i}{I_1}\right)$

dans laquelle Σp_i est la somme des mesures de taille pour toutes les zones de la population dans lesquelles l'échantillon d'EMO a été sélectionné. Ce paramètre demeure inchangé pour l'ETE avec les mêmes mesures de taille p_i.

Nous l'avons dit, on peut considérer g_1 comme le ratio des probabilités d'échantillonnage (f/f_1), qui est une constante pour toutes les unités, même dans un échantillon avec PPT existant. En effet, aussi bien f_1 que f sont définis, pour toutes les zones, proportionnellement à la taille p_i, qui s'annule dans le ratio ci-dessus.

Les conditions sont remplies: les zones de sondage des deux enquêtes ont la même structure et ne diffèrent que par le nombre des zones sélectionnés[16].

4.3. La distribution de l'échantillon et les domaines de rapport

4.3.1. Réduire le nombre des domaines de rapport

Mis à part les différences dans les conditions requises en matière de taille d'échantillon et d'échantillonnage par grappe, l'ETE peut aussi différer de l'EMO «parente» par les conditions relatives à la répartition et à la stratification de l'échantillon.

[16] Il convient de souligner que si le sous-échantillonnage à partir d'un échantillon avec PPT existant est très simple (la nature PPT restant inchangée avec un sous-échantillonnage à taux constant), le processus d'*ajout* d'unités additionnelles à un échantillon avec PPT est beaucoup plus complexe, et les probabilités résultantes des unités difficiles, sinon impossibles, à calculer. Nous supposons toujours que le nombre de zones requis pour l'ETE ne dépasse dans aucune strate d'échantillonnage le nombre déjà disponible dans l'échantillon d'EMO.

Par exemple, l'EMO peut être répartie de manière disproportionnée en vue de produire des estimations sous-nationales avec suréchantillonnage des petites régions ou d'autres domaines redditionnels, alors que cela ne sera pas exigé pour une ETE quand elle sera basée sur un plus petit échantillon visant essentiellement à produire des estimations au niveau national ou des répartitions limitées à quelques grands domaines.

Dans tous les cas, nous pouvons généralement attendre de l'ETE qu'elle porte sur un plus petit nombre de domaines redditionnels qu'une enquête plus importante comme l'EMO; de plus, les taux d'échantillonnage de l'échantillon d'ETE sont souvent plus uniformes. Normalement, les domaines redditionnels de l'ETE sont des groupements des domaines de rapports de l'EMO – par ex. les grandes régions plutôt que les provinces ou les districts[17].

Telle est la situation pour laquelle les procédures de sous-échantillonnage seront décrites ci-dessous.

Bien entendu, il peut arriver, à l'inverse, que l'ETE ait besoin d'un échantillon d'une grandeur disproportionnée pour des sous-populations d'intérêt particulier, ce suréchantillonnage dépassant les limites des domaines de l'EMO. On peut aussi estimer que, du point de vue d'une répartition optimale, les domaines qui comportent des niveaux plus élevés de travail des enfants devraient être échantillonnés à des taux plus élevés. Toutefois, cette répartition disproportionnée à des fins d'optimisation n'est pas justifiée dans la plupart des situations, eu égard aux gains d'efficience très modestes qu'elle apportera probablement, comme nous l'indiquons dans la section 4.1.1.

L'apparition de différences sensibles dans la stratification requise est peu probable. Certains critères de stratification communs, tels que la localisation géographique et le type de lieu (urbain-rural, degré d'urbanisation, etc.), sont couramment utilisés dans presque toutes les enquêtes auprès de chaque ménage. Dans de nombreuses situations, ce sont pratiquement les seuls critères dont on dispose pour la stratification. Quand ils sont disponibles, les critères de stratification additionnels (tels que l'appartenance ethnique, l'activité prédominante, le taux d'alphabétisation, le niveau moyen des revenus, etc., pour la zone de sondage) qui seront probablement utiles tendent aussi à être semblables pour des enquêtes sociales différentes. L'ETE et l'EMO seront probablement plus semblables encore en termes de stratification à cause du partage ou de la similitude de leurs thèmes. Les seules différences qui peuvent apparaître dans les conditions de la stratification sont en relation avec les différences dans les conditions de distribution (répartition) de l'échantillon entre les deux enquêtes.

Dans tous les cas, les différences dans les conditions de distribution de l'échantillon n'imposent pas *par elles-mêmes* aux deux enquêtes d'être stratifiées différemment. Il est possible de maintenir une séparation entre les conditions de répartition de l'échantillon (qui déterminent les probabilités de sélection requises) et les aspects de stratification et de sélection des échantillons (qui déterminent la mise en œuvre des procédures réelles de sélection). La méthode de sélection décrite dans la section 4.4 peut être utile à cet effet.

[17] Il est peu probable que la situation des domaines d'EMO, plus détaillés, recoupe les limites des domaines d'ETE, plus agrégés.

4.3.2. La procédure de sous-échantillonnage en cas de groupement de domaines d'EMO

Unités sélectionnées avec probabilité constante

Ici encore, il sera commode pour illustrer notre exposé de commencer par l'examen d'un cas simple.

Supposons qu'un certain nombre de domaines (j) de l'EMO soient groupés pour constituer un seul domaine de rapport pour l'ETE. L'échantillon d'EMO a été réparti de manière disproportionnée entre ces domaines. Appelons a_j le nombre de zones sélectionnées parmi A_j zones du domaine j avec une probabilité uniforme $f_j=a_j/A_j$. L'objectif est d'obtenir un échantillon d'ETE de *a* zones, sélectionnées à un taux uniforme $f=a/A$, avec $A=\Sigma A_j$, la somme des valeurs Aj, réuni à partir de tous les domaines d'EMO du groupe constitué.

Ici encore, le taux de sous-échantillonnage requis est donné par le ratio nouveaux taux d'échantillonnage du domaine concerné et taux d'échantillonnage existants de ce même domaine:

$$g_j=f/f_j=(a/A)/(a_j/A_j)=(a'_j/a_j)$$

où $a'_j=a\bullet(A_j/A)$ est le nombre moyen de zones de sondage de l'ETE dans le domaine j de l'EMO.

Nous supposons que dans chaque domaine l'échantillon d'ETE est le même que celui de l'EMO, ou un sous-échantillon de celui-ci:

$$a_j'\leq a_j, \text{ i.e. } g_j\leq 1.$$

Il s'agit là d'une hypothèse réaliste[18].

L'intervalle de sous-échantillonnage requis dans le passage des zones de sondage de l'EMO à celles de l'ETE est

$$k_j=1/g_j=(a_j/A_j)/(a/A)=(a_j/\ a'_j).$$

Le sous-échantillonnage requis peut donc être effectué séparément pour chaque domaine – en fixant le nombre requis à sélectionner a'_j et en appliquant l'intervalle de sélection k_j, qui diffère généralement d'un domaine à un autre.

Une procédure plus commode consiste à inscrire un domaine après un autre sur une seule liste, et à sélectionner un seul échantillon systématique de *a* unités. Le problème posé par les différents intervalles de sélection k_j à appliquer aux différentes parties de la liste peut être évité en utilisant l'astuce qui consiste à «réajuster les mesures de taille» des unités selon la méthode décrite dans la section 4.4. Ce réajustement s'effectue de telle sorte que l'application d'un intervalle d'échantillonnage commun à tous les domaines produit l'échantillon d'ETE requis.

[18] S'il apparaît que g_j excède 1,0 pour un domaine particulier, une option consiste à retenir simplement toutes les zones de l'EMO existantes pour l'ETE dans ce domaine, et à retirer ensuite le domaine des calculs décrits ici.

Sous-échantillonnage d'un échantillon avec PPT

Supposons maintenant que les zones d'EMO aient été sélectionnées avec une probabilité proportionnelle à une certaine mesure de taille, p_i.

Pour un domaine donné j, appelons P_j la somme des mesures de taille de toutes les unités de la *population* du domaine (et non pas seulement des zones de sondage de l'EMO), et a_j le nombre des zones sélectionnées dans l'EMO.

Nous supposons que pour l'ETE, un échantillon avec PPT de taille a est requis pour les zones sélectionnées, avec la *même* mesure de taille p_i. Les équations de sélection pour la zone i sont:

EMO, domaine j: $f^{(i)}_j=(a_j/P_j)\bullet p_i$

ETE, tous domaines: $f^{(i)}=(a/P)\bullet p_i$

où $P=\Sigma P_j$, la somme des valeurs Pj, portent sur tous les domaines, et où l'indice (i) indique que la probabilité de sélection des zones dans l'échantillon avec PPT varie en fonction des mesures de taille unitaires. Le taux de sous-échantillonnage requis pour le passage de l'EMO à l'ETE pour le domaine j est

$$g_j=f^{(i)}/f^{(i)}_j=(a/P)/(a_j/P_j)\ (a_j'/a_j) \text{ où } a_j'=a\bullet(P_j/P)$$

avec l'intervalle de sélection systématique requis:

$$k_j=(a_j/P_j)/(a/P)=(a_j/\ a_j')$$

Ce taux ne dépend ni de l'unité ni de sa taille, puisque p_i s'annule dans ce qui précède. La procédure de sous-échantillonnage est donc exactement la même que pour le cas de probabilité constante examiné ci-dessus: au sein de chaque domaine, des unités sont sélectionnées à partir de l'échantillon existant à un taux *constant* défini ci-dessus, quelles que soient leurs mesures de taille, à condition que celles-ci soient les mêmes pour les deux échantillons.

La méthode de réajustement des mesures de taille décrite ci-dessous peut être utilisée pour effectuer la sélection en une seule opération pour tous les domaines d'EMO regroupés.

4.4. Réajuster l'échelle des mesures de taille pour faciliter la sélection de l'échantillon

Cette technique est basée sur le *réajustement des mesures de taille* utilisées pour l'échantillonnage afin de traiter les variations requises des taux d'échantillonnage et des tailles d'échantillon entre les différentes parties de la liste. En fait, son utilité dépasse largement la question spécifique du sous-échantillonnage de l'EMO à l'ETE examinée ici.

Supposons qu'un échantillon systématique doit être sélectionné sur une liste d'unités. Celle-ci est divisée en un certain nombre de parties ou domaines, et les intervalles de sélection requis diffèrent d'une partie de la liste à une autre.

Le procédé s'applique à des situations diverses. Par exemple, l'objectif peut être de sélectionner un échantillon des unités avec probabilité constante ou avec PPT. Il peut également porter sur le sous-échantillonnage des unités à partir d'un échantillon existant, lui-même avec une probabilité constante ou variable. Ces détails n'affectent pas la procédure décrite ci-dessous, dans la mesure où ils sont incorporés dans la définition de l'intervalle d'échantillonnage I_j.

Une des options consiste à sélectionner séparément pour chaque partie j, de la manière habituelle, un échantillon systématique en utilisant l'intervalle de sélection I_j applicable pour cette partie.

Une procédure alternative plus simple comporte les démarches suivantes[19].

Soit $a=\Sigma a_j$ le nombre total d'unités à sélectionner sur la liste combinée, et $P'=\Sigma P'_j$ la somme pour l'ensemble des domaines combinés des mesures de taille modifiées selon la définition ci-dessus.

Appliquer un intervalle d'échantillonnage uniforme $I=P'/a$ à la liste avec des mesures de taille des unités modifiées équivaut à appliquer des intervalles de sélection I_j différents aux unités ayant conservé leurs mesures de taille initiales. Avec des mesures de taille réajustées comme ci-dessus, l'intervalle de sélection à appliquer devient partout uniforme pour tous les domaines. Ceux-ci peuvent donc être placés à la suite l'un de l'autre, sur une liste unique, en vue d'une sélection d'échantillon à intervalle uniforme.

Si dans un ensemble de domaines, P_j soit la somme des mesures de taille des unités (UPE) de la *population* du domaine j, et si l'on doit sélectionner dans ce domaine a_j unités avec PPT. Cela nous donne un intervalle de sélection $I_j=P_j/a_j$.

Et si que toutes les mesures de taille des unités du domaine j soient multipliées par (I/I_j), où I est une constante arbitraire. La nouvelle somme des mesures de taille pour ce domaine devient $P'_j=P_j\bullet(I/I_j)$.

Pour sélectionner dans ce domaine le même nombre d'unités a_j qu'auparavant, mais en utilisant les mesures de taille réajustées, l'intervalle de sélection requis est

$$[P_j\bullet(I/I_j)/a_j]=I\bullet[(P_j/I_j)/(P_j/I_j)]=I$$

égal à la constante arbitrairement choisie I pour chaque domaine.

L'échantillon à probabilité constante n'est qu'un cas particulier de ce qui précède, avec les mesures de taille initiales des unités p_i=constante=1 par exemple. Dans ce cas,

[19] Les deux procédures ne sont pas identiques. La solution plus simple décrite ici permet quelques variations aléatoires dans le nombre exact des sélections qui tombent dans tel ou tel domaine, mais elle évite d'avoir à répartir séparément à chacun des domaines initiaux des nombres de sélections arrondis prédéterminés – chose qui peut être inutile et aussi malcommode quand un grand nombre de petits domaines sont concernés, notamment quand le nombre de sélections moyen par domaine est limité. La procédure évite d'avoir à arrondir le nombre des unités à sélectionner.

l'échantillonnage à effectuer est maintenant traité comme PPT, avec chaque unité en j ayant la même mesure de taille (I/I_j).

Cette technique peut être très commode quand un nombre important de parties est impliqué, chacune exigeant un intervalle d'échantillonnage différent, mais que le procédé de l'échantillonnage n'exige pas de ces parties qu'elles forment des strates explicites.

Cette technique commode peut aussi être appliquée séparément à différents stades d'un échantillon à degrés multiples, simplement en réajustant indépendamment, de manière appropriée, les mesures de taille à chaque stade de la sélection. Par exemple, considérons un échantillon autopondéré en deux stades dans lequel les unités sont sélectionnées de la manière habituelle, en utilisant la PPT au stade 1 et la PPT inverse au stade 2:

$$f1_i = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}; \quad f2_i = \frac{b}{p_i}; \quad f = f1_i * f2_i = \frac{b}{I}$$

Supposons maintenant que dans une partie de la population, des taux d'échantillonnage différents soient requis, d'un facteur k_1 au stade des zones et de k_2 au stade final. Pour que cette partie soit échantillonnée différemment, les équations de sélection deviennent:

$$\mathrm{f1_i} = \left(\frac{\mathrm{k_1 * a}}{\Sigma \mathrm{p_i}}\right).\mathrm{p_i}; \quad \mathrm{f2_i} = \frac{(\mathrm{k_2 * b})}{\mathrm{p_i}}; \quad \mathrm{f} = (\mathrm{k_1.k_2}) * \mathrm{f}$$

Bien entendu, on peut y parvenir en plaçant la partie à échantillonner à un taux différent dans une strate séparée et en effectuant le processus de sélection séparément dans les deux parties obtenues. Cela implique d'utiliser des intervalles d'échantillonnage différents pour la sélection systématique d'unités dans le domaine concerné.

On obtient toutefois le même résultat en maintenant inchangés les intervalles d'échantillonnage pour une sélection systématique, mais en augmentant la mesure de taille des zones concernées selon le facteur k_1 pour la sélection du premier stade (donnant la nouvelle mesure de taille $P'_i = k_i \cdot p_i$), et selon le facteur $1/k_2$ pour la sélection du second stade au sein des zones de sondage dans le domaine particulier concerné (donnant la nouvelle mesure de taille $P''_i = P_i/k_2$). Les mêmes équations de sélection se lisent:

$$f1_i = \frac{(k_1 * p_i)}{I} = \left(\frac{P'_i}{I}\right); \; f2_i = \frac{b}{(p_i/k_2)} = \left(\frac{b}{P''_i}\right); \; f = f1_i * f2_i = (k_1.k_2) * f$$

Après cet ajustement, il n'est pas nécessaire de sélectionner l'échantillon séparément du domaine concerné. Avec l'échantillonnage systématique, par exemple, le même intervalle de sélection commun «I» peut être appliqué à la liste tout entière pour la sélection des zones. Les mesures de taille multipliées par le facteur k_1 répartissent automatiquement l'échantillon de zones selon les conditions requises au domaine concerné. De même, les prélèvements d'échantillon par zone sont ajustés

automatiquement quand on utilise l'intervalle initial de sélection *b* pour un «tirage d'échantillon cible», avec les mesures de taille des zones du domaine concerné divisées par le facteur k_2 pour ajuster le nombre d'unités finales à sélectionner dans chaque zone de sondage. On notera que pour le(s) même(s) zone(s), l'échelle des mesures de taille a été ajustée différemment aux deux stades (en fonction de k_1 et de $1/k_2$ respectivement).

Une illustration

La procédure est basée simplement sur l'observation selon laquelle, dans l'échantillonnage avec PPT, le nombre des unités sélectionnées

$$a=\Sigma p_i/I$$

demeure inchangé si les mesures de taille comme l'intervalle de sélection sont multipliés par la même constante arbitraire. En choisissant la constante appropriée pour chaque domaine à échantillonner selon un taux différent, nous pouvons faire en sorte que l'intervalle de sélection requis soit le même dans tous les domaines.

Supposons qu'il existe deux strates. La première consiste en $A_1=100$ zones, dont $a_1=10$ doivent être sélectionnées avec une probabilité constante, autrement dit $f_1=a_1/A_1=1/10$ et l'intervalle de sélection $I_1=10$. Dans le second domaine, les chiffres sont respectivement $A_2=40$, $a_2=8$, d'où $f_2=1/5$, $I_2=5$.

Nous souhaitons sélectionner l'échantillon de manière systématique, mais en menant pour plus de commodité une seule opération d'échantillonnage systématique. Remplaçons comme suit les sélections à probabilité constante par un échantillonnage systématique avec PPT équivalent: chaque unité de la strate 1 reçoit une mesure de taille de 1 (il n'y a donc pas de changement); dans la strate 2, nous attribuons une mesure de taille de 2 à chaque zone, et par conséquent nous doublons aussi l'intervalle de sélection à appliquer à 5*2=10, qui est maintenant identique à celui de la strate 1.

	Strate 1	Strate 2	Total
Nombre d'unités dans la population	100	40	140
Mesures de taille attribuées à chaque unité	1	2	
Mesure de la taille totale	100	80	180
Nombre d'unités sélectionnées moyen (avec I=10)	10	8	18

Le «prix à payer» pour avoir réuni les deux strates sur une seule liste est que le nombre réel d'unités sélectionnées pour une strate particulière peut varier légèrement de manière aléatoire (bien que les probabilités de sélection demeurent strictement inchangées). Mais cette procédure peut être commode quand on a affaire à de nombreux domaines distincts.

4.5. Traiter de «très grandes» zones dans le sous-échantillonnage

Des démarches complémentaires sont incluses dans la procédure de sous-échantillonnage quand on a affaire à des zones de l'échantillon initial qui ont été sélectionnées selon des procédures spéciales parce qu'elle étaient considérées comme trop «petites» ou trop «grandes» pour la procédure avec PPT normale. Elles ont été décrites dans les sections 3.9-10.

En ce qui concerne le sous-échantillonnage examiné dans ce chapitre, traiter des zones classifiées comme «petites» au sens évoqué ci-dessus est un peu plus complexe, car cela fait intervenir, dans un plan à degrés multiples, des questions concernant la sélection des ménages au sein des zones de sondage. Nous y reviendrons dans la section 4.7.

Considérons d'abord les très grandes unités. Comme nous l'avons définie dans la section 3.9, on considère comme «très grande», dans le contexte de l'échantillonnage avec PPT, une unité dont la mesure de taille excède l'intervalle d'échantillonnage, c'est-à-dire que $p_i > I$. Supposons que dans l'EMO, ces zones aient été traitées avec l'option recommandée (3) de la section 3.9. Cela implique de considérer la grande unité comme automatiquement sélectionnée (ces unités sont habituellement appelées «unités autoreprésentées»). Si les unités finales qu'elle contient sont ensuite sélectionnées avec la probabilité de sélection globale requise, appelé f, cela donnera un échantillon autopondéré avec

$f1_i = 1; \quad f2_i = f$ pour l'EMO

Dans le but de sélectionner un sous-échantillon (1/k) d'unités sectorielles de l'EMO pour l'ETE, il convient de distinguer deux groupes de ces zones d'EMO «autoreprésentées». Ci-dessous, p_i est la mesure de taille unitaire et I l'intervalle de sélection avec PPT.

Groupe 1: $p_i > k * I$

Ce sont les plus grandes unités. Elles doivent toutes être retenues dans l'ETE, avec une probabilité $=1$, comme dans l'EMO. Au dernier stade, les unités finales peuvent être sélectionnées avec le taux de sélection global requis pour l'ETE.

Group 2: $I < p_i \leq k * I$

Toutes ces grandes unités ne sont pas sélectionnées automatiquement dans l'ETE, bien qu'elles l'aient été dans l'EMO. Pour ces unités, un échantillon de zones approprié avec PPT peut être sélectionné pour l'ETE avec

$$f1'_i = \frac{p_i}{(k * I)} \leq 1$$

Pour constituer l'échantillon final autopondéré, par exemple, l'échantillonnage final au sein des zones sélectionnées se fait à un taux inversement proportionnel à p_i, avec les constantes appropriées pour obtenir le taux de sélection total requis pour l'ETE.

Une illustration

Le tableau 4.1 présente la liste des «grandes» zones définies dans le tableau 3.3. Les zones numérotées 93-100 sont «grandes» selon la perspective de l'EMO parce que leurs mesures de taille excèdent 500, soit l'intervalle d'échantillonnage pour la sélection de zones avec PPT prévue pour l'EMO.

Ces zones ont été traitées comme si la mesure de taille de chacun était de 500, si bien que la zone a été sélectionnée dans l'EMO avec une probabilité de 1,0. (Au stade suivant, si la sélection des ménages s'effectue avec une PPT inverse, la zone apportera davantage de ménages à l'échantillon en proportion de sa taille réelle.)

Supposons que l'intervalle de sélection pour l'ETE soit de 700. Les «grandes» zones de l'EMO se trouvent alors divisées en deux groupes. Les zones numérotées 93-96 cessent d'être «grandes» en ce qui concerne l'ETE – leurs mesures de taille sont plus petites que l'intervalle d'échantillonnage de l'ETE, soit 700. À partir de l'EMO, ces zones doivent être sélectionnées avec PPT, c'est-à-dire avec une probabilité égale à la mesure de taille des unités divisée par 700 (col. [7] du tableau 4.1).

Le second groupe, celui des zones numérotées 97-100, reste «grande» au regard de l'intervalle d'échantillonnage de l'ETE. Ces zones sont retenues dans l'échantillon de l'ETE avec une probabilité de 1,0. (Comme dans le cas de l'EMO, si la sélection des ménages au stade suivant s'effectue avec une PPT inverse, une «grande» zone apportera davantage de ménages à l'échantillon en proportion de sa taille réelle.)

Tableau 4.1: Illustration: sous-échantillonnage de «grandes» zones de l'EMO à l'ETE

«Premier échantillon»: par exemple pour une EMO.

Paramètres:

a = 100 [nombre de zones dans l'échantillon]
b = 20 [nombre moyen d'unités sélectionnées/zone d'échantillonnage]
I = 500 [intervalle pour la sélection systématique des zones]
n = 2,000 [=a*b, taille moyenne d'échantillon]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
Fréquence cumulative du nombre de zones		Mesure de la taille de la zone	MoS modifié pour EET	Probabilité de sélection d'une zone	MoS modifié pour ETE	Probabilité de sélection d'une zone
dans la population	échantillon	pi	pi modifié	f1 (%)	pi modifié	f1' (%)
Très grandes zones						
3,0	1	8	20	4,00	20	1,09
6,0	2	11	20	4,00	20	1,61
9,0	3	12	20	4,00	20	1,72
12,1	4	16	20	4,00	20	2,25
15,1	5	17	20	4,00	20	2,42
18,1	6	18	20	4,00	20	2,50
21,1	7	18	20	4,00	20	2,63
24,1	8	20	20	4,00	20	2,85

Zones normales						
…..						
99,0	92	425	425	85,06	425	60,76
…..						
Très petites zones						
99,2	93	527	500	100,00	527	75,24
99,3	94	566	500	100,00	566	80,83
99,4	95	659	500	100,00	659	94,08
99,5	96	681	500	100,00	681	97,34
99,6	97	719	500	100,00	700	100,00
99,8	98	898	500	100,00	700	100,00
99,9	99	908	500	100,00	700	100,00
100,0	100	1076	500	100,00	700	100,00

4.6. De l'EMO à l'ETE: sous-échantillonnage des unités finales au sein des zones de sondage

Il n'est en aucune façon automatiquement nécessaire ou utile de confiner l'ETE au même échantillon, ou à un sous-échantillon, d'unités finales (ménages, personnes, enfants) incluses dans l'EMO, même quand les deux enquêtes partagent un ensemble commun de zones de sondage. Il existe une série de choix possibles en fonction des circonstances et des conditions imposées.

1. À une extrémité, les zones de sondage communes EMO-ETE peuvent être soumises à un nouveau listage pour obtenir une base actualisée pour l'ETE, et un échantillon entièrement nouveau des unités finales sélectionnées. Toutefois, le listage tend à être une opération coûteuse, et le coût d'un relistage ne peut se justifier que si les deux enquêtes sont séparées par un long intervalle de temps, comme une année ou plus dans de nombreuses circonstances. D'un autre côté, l'utilisation de listes anciennes peut introduire des erreurs de couverture, dans la mesure où des unités nouvellement créées (comme des logements ou des ménages nouveaux) ne sont pas représentées. Par ailleurs, les listes anciennes rendent plus difficile l'identification des unités sélectionnées sur le terrain. La présence sur une liste d'unités qui n'existent plus peut rendre difficile l'identification des véritables non-répondants; elle tend aussi à réduire l'efficience de l'opération d'échantillonnage et le contrôle exercé sur celle-ci. La pérennité des listes dépend du type des unités listées: les adresses ou les unités de logement tendent à être plus durables que les ménages, et les ménages plus durables que les personnes.

2. Quand les mêmes listes sont utilisées, les deux échantillons (EMO et ETE) peuvent encore être sélectionnés indépendamment, ou du moins avec un recouvrement réduit au minimum. Cela est souhaitable quand la lassitude des répondants est prise en compte, ou quand le premier échantillon est sujet à des taux élevés de non-réponses. Ce dernier problème peut être sérieux si l'échantillon d'ETE est basé sur une enquête «lourde» plutôt que sur une EMO typique. Un échantillonnage

indépendant ou au moins complémentaire est également requis quand la première enquête n'est pas en mesure de produire un échantillon assez grand pour l'ETE.

3. L'option suivante consiste à baser l'ETE sur la totalité ou sur un sous-ensemble des unités finales incluses dans le premier échantillon. Le résultat dépend du type et des caractéristiques des unités incluses dans le sous-échantillonnage. Le plus simple est souvent de sélectionner un sous-échantillon d'adresses. Vient ensuite l'utilisation des ménages comme unités de sous-échantillonnage. La démarche est plus simple si les ménages de la première enquête font l'objet d'un sous-échantillonnage sans aucune référence à leurs caractéristiques particulières.

4. Il arrive toutefois que des informations concernant diverses caractéristiques des ménages soient évoquées à des fins de stratification ou pour l'application de taux d'échantillonnage différents, ce qui implique la collecte de ces informations dans le cadre de l'EMO et leur préservation et leur transfert à l'ETE. Cela peut être coûteux et encombrant, et dans bien des cas assez peu efficace pour améliorer l'efficience de l'échantillon résultant.

5. Il est également courant d'exclure de la sélection certains types de ménages, comme ceux dont on constate qu'ils ne comprennent pas d'enfants relevant de l'ETE. Cette disposition peut améliorer le contrôle de la taille d'échantillon de l'ETE, ainsi que l'efficience de son travail de terrain. Elle suppose que la situation des ménages concernant les critères d'exclusion n'a pas changé dans l'intervalle qui sépare les deux études. Un autre problème est, qu'entre les deux enquêtes, il ne faut pas seulement transférer la liste d'adresses, mais aussi des informations complémentaires sur les ménages.

6. L'utilisation des enfants identifiés au cours de la première enquête comme unités d'échantillonnage pour l'ETE est également une option. Ce choix impose toutefois des obligations: les listes d'enfants identifiés doivent être préparées et transférées aux opérations de l'ETE pour échantillonnage, puis les enfants sélectionnés doivent être identifiés dans le cadre des activités de terrain. Des erreurs d'identification des enfants peuvent facilement survenir. En règle générale, cette option ne doit être adoptée que si l'intervalle entre les deux enquêtes est très court et ne dépasse pas quelques semaines.

7. À un extrême se situe la procédure selon laquelle les informations concernant diverses caractéristiques des enfants, comme leur situation sur le plan de l'éducation et/ou des activités, sont utilisées à des fins de stratification ou pour l'application de taux d'échantillonnage différents. Cette procédure peut paraître attractive quand la taille d'échantillon de l'ETE est très petite et que sa structure doit être étroitement contrôlée. Toutefois, *il s'agit généralement d'une procédure exigeante et coûteuse, exposée aux erreurs d'exécution*. Elle ne doit être utilisée que lorsque l'ETE est proche de remplir les conditions qui en feraient un «module» de la première enquête, c'est-à-dire être quasi simultanée ou proche dans le temps, et s'appuyer dans une mesure substantielle sur l'information fondamentale collectée au cours de la première enquête.

4.7. Traiter de «très petites» zones dans le sous-échantillonnage

La procédure de sous-échantillonnage devient plus complexe quand on traite des unités de très petite taille. Il faut être attentif à ce que des probabilités de sélection correctes soient appliquées dans l'ETE.

Dans un plan à deux degrés, la procédure de traitement de très petites zones dans l'échantillonnage de l'EMO à l'ETE dépend des modalités de l'échantillonnage de l'EMO aux deux stades. Pour décrire la procédure de sous-échantillonnage, nous supposerons le plan d'autopondération suivant. À l'intérieur de chaque zone de sondage de l'EMO sélectionné avec une probabilité proportionnelle à P_i, les ménages sont supposés avoir été sélectionnés avec une probabilité inversement proportionnelle à p_i, avec une moyenne de b=n/a ménages par zone de sondage[20].

Pour l'ETE, le sous-échantillonnage à partir de l'EMO comporte deux stades: la sélection de 1 sur k zones de sondage, a"=a/k puis la sélection d'un nombre moyen b' (moyenne ou cible) d'unités finales par zone inclus dans l'ETE, donnant une taille d'échantillon cible de l'ETE de n'=a'*b'[21].

Les équations complètes de sélection (pour un plan autopondéré) dans les deux enquêtes sont présentées dans le tableau 4.2.

Tableau 4.2: Équations de sélection supposées de l'EMO et de l'ETE décrivant la procédure de sous-échantillonnage de l'EMO à l'ETE

	(Dernier) stade sectoriel	Stade final	Probabilité de sélection globale
EMO	$f1_i = \frac{p_i}{I}$	$f2_i = \frac{b}{p_i}$	$f = f1_i * f2_i = \frac{b}{I}$
ETE[22]	$f1'_i = \frac{p_i}{I'}, I' = k \cdot I$	$f2'_i = \frac{b'}{p_i}$	$f' = f1'_i * f2'_i = \frac{b'}{I'} = \left(\frac{1}{k} \cdot \frac{b'}{b}\right) * f$

[20] D'autres options sont également possibles, comme le tirage fixe de b ménages par zone, ou un compromis entre les versions autopondérée et à tirage fixe du plan, comme nous l'avons exposé dans le chapitre 3. Nous avons choisi ici la version autopondérée pour simplifier l'exposé.

[21] On notera que les deux échantillons ne se superposent pas nécessairement dans le cas des unités finales, bien qu'ils le fassent souvent et que la condition b'=b est réalisée. De plus, pour l'EMO comme pour l'ETE, ces paramètres peuvent différer d'un «domaine d'échantillonnage» ou d'une strate à l'autre. Il suffit ici de décrire la procédure appliquée à un seul de ces domaines.

[22] On notera que dans les équations ci-dessus, b' doit être défini à la même échelle que p_i ou b, telle que la population totale ou le nombre des ménages. Si b' est à une autre échelle, telle que le nombre des enfants, alors p_i et b, dans les équations concernant l'ETE, sont supposées réajustées à la même unité par un facteur constant.

Comme nous l'avons exposé dans la section 3.10, on qualifie de «très petite», dans le contexte de l'échantillonnage avec PPT, une unité dont la mesure de taille est plus petite que le prélèvement d'échantillon requis, autrement dit $p_i < b$ dans l'EMO. Ce qu'il convient de faire pour le sous-échantillonnage de l'EMO à l'ETE du traitement de ces zones au cours de la sélection de l'échantillon d'EMO lui-même. Plusieurs méthodes permettant de traiter ce problème ont été présentées dans la section 3.10. De plus, nous devons également examiner la relation de la taille p_i avec le prélèvement d'échantillon b' requis dans l'ETE.

Il est important de souligner que si nous considérons un échantillon d'ETE dont la taille globale est plus petite par comparaison avec l'EMO (n'<n) et qui comporte un nombre inférieur de zones de sondage (a'<a), il ne s'ensuit pas nécessairement que le tirage d'échantillon par zone (b') dans l'ETE doit être plus petit que le tirage (b) dans l'EMO. Par exemple, à l'intérieur des zones de sondage ordinaire, l'ETE peut utiliser les mêmes échantillons des unités finales que l'EMO (b'=b), comme ce serait le cas si la première enquête formait un module de la seconde; ou bien l'ETE peut utiliser un sous-échantillon des unités des grappes communes (b'<b); ou encore les deux enquêtes peuvent utiliser des échantillons d'unités finales entièrement différents, même dans les zones de sondage communes. Dans ce dernier cas, chacune des trois possibilités existe: b'<b, b'=b, ou b'>b. Cela tend à compliquer quelque peu le traitement adéquat des petites zones dans le sous-échantillonnage de l'EMO à l'ETE. Néanmoins, la question revêt une importance pratique et demande à être traitée correctement.

Option (3) de la section 3.10

Supposons que dans l'EMO de très petites zones aient été sélectionnées selon l'option mentionnée ci-dessus dans l'EMO. Cette option impliquait que tous les très petites zones, avec p_i recevaient une mesure de taille =b, et étaient donc sélectionnées avec une probabilité $f1_i$=(b/I) au stade sectoriel; toutes les unités finales d'une zone sélectionnée étaient incluses dans l'échantillon d'EMO ($f2_i$=1), donnant la probabilité globale requise f=(b/I) pour l'EMO.

Le tableau 4.3 présente tous les détails de la sélection de zones de l'EMO à l'ETE pour les très petites unités. Les procédures relatives aux zones «normales» et très grandes, qui ont été examinées dans les sections 3.9 et 4.5, sont également présentées pour permettre les comparaisons. La colonne «Taux de sous-échantillonnage des zones» indique la proportion de zones de sondage de l'EMO qui, en fonction de leur taille p_i, seront retenus dans l'ETE pour obtenir le taux de sélection total requis f'=f1'*f2'=b'/(k*I).

Pour effectuer cette sélection, une méthode simple consiste à attribuer aux zones de sondage de l'EMO des mesures de taille modifiées (différentes des tailles p_i initialement utilisées pour la sélection des zones d'EMO), telles qu'elles sont définies dans la dernière colonne du tableau 4.3. Nous pouvons ensuite choisir un sous-échantillon systématique des zones de sondage de l'EMO avec l'intervalle k pour obtenir l'échantillon sectoriel de l'ETE. Cette procédure a été exposée dans la section 4.4.

Option (5) de la section 3.10

Il est également utile de donner des détails pour cette option, puisque c'est celle qui est recommandée. Dans celle-ci, les très petites zones de l'EMO sont divisées en deux parties: (i) un certain nombre des plus petites zones, disons avec $p_i<b_0$, reçoivent une mesure de taille =0 et sont donc effectivement exclus de la base de sondage. Ces zones sont considérées comme trop petites pour être incluses dans un échantillon d'EMO. (ii) les zones restantes, soit les plus grandes des «très petites» zones, reçoivent chacune une mesure de taille b et sont traitées comme dans l'option (3).

On suppose que, pour compenser (i), la pondération de chacune des unités sélectionnées dans l'EMO a été augmentée par le facteur (b/p_i). Comme on l'a noté dans la section 3.10, pour cette compensation, la partie (ii) devrait inclure $A = \Sigma p_i / b$ des plus grandes unités de l'ensemble des «très petites», Σp_i étant la mesure de taille totale de toutes les unités de l'ensemble.

Avec les zones d'EMO sélectionnées avec l'option 5 comme indiqué ci-dessus, la seule différence par rapport à l'option 3 concerne les cas 5 et 6 du tableau 4.3.

Bien entendu, les zones avec $p_i<b_0$ n'apparaissent pas dans l'échantillon d'ETE, puisqu'elles ont déjà été exclues de l'EMO.

Pour les zones avec $p_i \geq b_0$, l'augmentation de pondérations dans l'échantillon d'EMO ($=b/p_i$) mentionnée ci-dessus est retenue pour toute zone sélectionnée pour l'ETE. Une solution de substitution plus simple consiste à modifier les deux dernières colonnes du tableau 4.3, pour les cas 5 et 6, par ce facteur (b/p_i), évitant ainsi la nécessité d'inclure explicitement cette pondération dans l'ETE. Cela a simplement pour effet de rendre le traitement du cas 6 identique à celui du cas 4, et de même que le traitement du cas 5 à celui du cas 3 – ce qui a été fait dans les deux dernières colonnes du tableau 4.3.

Tableau 4.3: Sous-échantillonnage des zones d'EMO pour l'ETE

	Condition	EMO		ETE		Taux de sous-échantillonnage des zones f'/f1	Mos[23]
		f1	f2	f1'	f2'		
Très grandes zones							
1	$k*I \le p_i$	1	$\frac{b}{I}$	1	$\frac{b'}{k*I}$	1	k
2	$I \le p_i < k*I$	1	$\frac{b}{I}$		$\frac{b'}{p_i}$	$\frac{1}{k}*\frac{p_i}{I}$	$\frac{p_i}{I}$
Zones normales (majorité des zones)							
3	$b, b' \le p_i < I$	$\frac{p_i}{I}$	$\frac{b}{p_i}$	$\frac{p_i}{k*I}$	$\frac{b'}{p_i}$	$\frac{1}{k}$	1
Très petites zones (sélectionnées dans l'EMO en utilisant l'option 3 de la section 3.10)							
4[24]	$b \le p_i < b'$	$\frac{p_i}{I}$	$\frac{b}{p_i}$	$\frac{b'}{k*I}$	1	$\frac{b'}{k*I}$	$\frac{b'}{p_i}$
5	$b' \le p_i < b$	$\frac{b}{I}$	1	$\frac{p_i}{k*I}$	$\frac{b'}{p_i}$	$\frac{1}{k}*\frac{p_i}{b}$	$\frac{p_i}{b}$
6	$p_i \le b, b'$	$\frac{b}{I}$	1	$\frac{b'}{k*I}$	1	$\frac{1}{k}*\frac{b'}{b}$	$\frac{b'}{b}$
Très petites zones (différence par rapport à ci-dessus en cas de sélection des zones dans l'EMO en utilisant l'option 5 de la section 3.10)							
5	$b_0, b' \le p_i < b$	Comme dans le cas 5 ci-dessus				$\frac{1}{k}$	1
6	$b_0 \le p_i \le b, b'$	Comme dans le cas 6 ci-dessus				$\frac{1}{k}*\frac{b'}{p_i}$	$\frac{b'}{p_i}$
7	$p_i < b_0$	Zones non incluses dans l'échantillon d'EMO ou d'ETE				-	0

[23] Mesure de taille modifiée assignée aux zones d'EMO pour un sous-échantillonnage avec intervalle k, en vue d'obtenir de manière simple les zones de sondage d'ETE. Ainsi, par exemple, chaque zone d'EMO de la ligne 1 reçoit une mesure de taille k, si bien que la sélection avec l'intervalle k retient simplement l'ensemble de ces zones dans l'échantillon d'ETE – ce qui était requis. De même, dans la ligne 2, chaque zone d'EMO reçoit une mesure de taille (p_i/I), si bien que la sélection avec le même intervalle k donne le taux de sous-échantillonnage requis (f1'/f1) spécifié dans la colonne précédente.

[24] On notera qu'un seul des deux cas 4 et 5 peut s'appliquer dans une situation particulière.

Chapitre 5
L'enquête sur les enfants travailleurs (EET)

5.1. L'approche de l'enquête

Selon la définition du chapitre 2 (section 2.1.2), nous utilisons l'expression *enquête sur les enfants travailleurs* (EET) par référence spécifique à une enquête ou à une composante d'enquête conçue avec l'objectif de déterminer les conditions et les conséquences du travail des enfants, à distinguer de son ampleur parmi l'ensemble des enfants (ce qui est l'objectif de l'enquête ou de la composante d'enquête appelée ETE).

Comme nous l'avons indiqué, l'objectif primordial de l'EET est d'explorer les circonstances, les caractéristiques et les conséquences du travail des enfants: les types d'enfants qui sont engagés dans des activités liées au travail, le type de travail effectué par les enfants, les circonstances et les conditions dans lesquelles les enfants travaillent, les effets du travail sur leur éducation, leur santé, leur développement physique et moral, et ainsi de suite. Ces objectifs peuvent aussi inclure des recherches sur les causes et les conséquences immédiates de l'entrée des enfants dans le travail. La base de population appropriée de l'EET est par conséquent la *population d'enfants travailleurs*.

Une question cruciale d'importance pratique est de savoir si les objectifs distincts de l'ETE et de l'EET peuvent être atteints de manière satisfaisante grâce à une unique enquête intégrée, ou s'il vaut mieux les organiser comme deux opérations séparées – et néanmoins liées. Une question connexe, quoique distincte, se pose également: les deux composantes peuvent-elles être basées sur le même échantillon de ménages, ou bien l'EET doit-elle être un sous-échantillon de l'ETE, de taille plus réduite et éventuellement de structure différente ?

Il peut sans aucun doute être plus commode et économique de couvrir les deux composantes en une seule opération, et la forme prédominante dans certains pays a été celle de l'opération ETE-EET intégrée. Si ce type d'arrangement peut être considéré comme plus économique, ce n'est malheureusement pas toujours le cas. Sur la base de l'étude d'enquêtes nationales passées sur le travail des enfants, nous considérons que les questions relatives aux différences entre les conditions fondamentales et statistiques, aux aspects pratiques tels que la charge imposée aux enquêtés et même au coût des enquêtes, et notamment à la qualité des données, n'ont pas toujours été prises en considération de manière approfondie dans les décisions relatives à la structure appropriée de l'enquête.

Une opération d'ETE-EET intégrée peut fort bien être l'option la plus adéquate dans certaines conditions, telles que celles-ci:

1. L'ETE n'exige pas un très large échantillon, ce qui est le cas quand elle n'a pas à produire des estimations détaillées de l'ampleur du travail des enfants pour un grand nombre de régions, de groupes de population, de secteurs d'activité ou d'autres types de domaines.

2. L'ETE est une enquête indépendante, si bien que son échantillon peut être conçu comme un compromis répondant à deux types de besoins d'information: estimer l'ampleur du travail des enfants avec la précision requise, d'une part, et explorer de manière détaillée les conditions et les conséquences du travail des enfants, d'autre part.

3. Le travail des enfants n'est pas trop hétérogène ou trop inégalement réparti pour être «saisi» de manière acceptable par un échantillon à finalité générale de la population d'enfants.

4. La taille d'échantillon de compromis obtenue n'est pas trop grande pour l'investigation approfondie que requiert habituellement la composante EET.

5. Mais, dans tous les cas, l'entretien intégré obtenu n'est pas assez lourd pour avoir un effet négatif sur la qualité (et notamment la complétude) de l'estimation de l'ampleur du travail des enfants, qui est l'objectif de la composante ETE. Il s'agit là d'un problème courant qui a été rencontré dans de nombreux autres types d'enquêtes ayant la même dualité d'objectifs.

Quand une ou plusieurs des conditions ci-dessus ne sont pas remplies, il est nécessaire d'envisager au moins la possibilité de séparer les composantes ETE et EET au niveau opérationnel.

Dans le cas de cette séparation, il sera généralement approprié d'envisager de baser la composante EET sur un sous-échantillon de l'ETE. Les objectifs du sous-échantillonnage seront à la fois de réduire la taille d'échantillon de l'EET et de faire que celle-ci soit plus concentrée et mieux ciblée, afin de refléter la répartition géographique inégale du travail des enfants.

Ce chapitre a pour objectif spécifique d'exposer en détail les techniques de ce sous-échantillonnage de l'ETE à l'EET. Nous nous préoccuperons avant tout des questions de conception d'un échantillon pour une enquête dans laquelle la population de base concernée est la *population d'enfants travailleurs*. Nous tenons pour acquis le fait que, comme nous l'avons évoqué dans d'autres chapitres de ce manuel, la population concernée est clairement définie sur la base de considérations substancielles et politiques, y compris celles qui sont liées à la nécessité de disposer de données internationalement comparables.

Il importe de préciser que le concept d'«enquête sur les enfants travailleurs» ne signifie aucunement que les unités finales dénombrées dans l'enquête seront seulement des enfants travailleurs. Au contraire, il sera habituellement nécessaire dans ce type d'enquête de dénombrer également les enfants qui ne sont pas engagés dans le travail, afin de disposer d'un groupe de contrôle permettant les comparaisons avec les caractéristiques et la situation de ceux qui sont astreints au travail des enfants. La signification du concept d'EET est que, si l'objectif est de déterminer les circonstances et les conséquences du travail des enfants plutôt que sa simple ampleur, alors la structure et la taille de l'échantillon devraient être essentiellement déterminées par la taille et la répartition de la *population d'enfants travailleurs*, plutôt que par la taille et la répartition de la population générale de tous les enfants. De plus, pour ces motifs comme pour d'autres raisons fondamentales et pratiques, il est

souvent souhaitable de *lier* de manière appropriée l'EET à l'ETE normale décrite au chapitre précédent, lorsqu'il existe une possibilité de réaliser cette liaison.

Enquêtes régulières auprès des ménages ou enquêtes ciblées

Il convient de noter les limites du champ des questions d'échantillonnage examinées dans ce chapitre: il traite de ce que l'on peut appeler les enquêtes «régulières» auprès des ménages sur les enfants travailleurs.

L'estimation de l'incidence et de la nature du travail des enfants dans des secteurs et des activités ciblés, et aussi de certains aspects des pires formes de travail des enfants, implique des particularités de conception dont certaines sont très différentes des type d'opérations d'EMO/ETE/EET, dont la base est plus large. À certaines fins et dans certaines circonstances, elles peuvent impliquer une collecte de données non basée sur les ménages, voire même des écarts par rapport aux principes d'échantillonnage probabiliste. Comme l'a indiqué le Comité consultatif externe du SIMPOC, les diverses techniques statistiques utilisées pour échantillonner des unités non standard doivent être développées et documentées. Il s'agit là d'une question majeure qui nécessite un traitement séparé et qui n'est pas évoquée dans ce manuel. Néanmoins, un grand nombre des techniques qui sont évoquées ici peuvent aussi être utiles dans la conception d'enquêtes plus spécialisés, ciblées ou sectorielles sur les enfants travailleurs.

La base de sondage de l'EET

Les procédures décrites ci-dessous pour la sélection des zones de sondage de l'EET demandent des informations sur le *nombre des enfants travailleurs* de chaque zone de la base dans laquelle sont sélectionnés les zones de sondage. Ce type d'information n'est habituellement pas disponible dans les bases de population à usage général, si bien qu'il est nécessaire de sélectionner l'EET comme sous-échantillon des zones pour lesquels cette information a été collectée, comme dans les zones d'échantillon d'une ETE de plus grande ampleur.

On peut s'attendre à trouver ici des considérations sur la conception des échantillons qui seront à bien des égards semblables à celles du chapitre précédent concernant l'enquête sur le travail des enfants (ETE). La différence essentielle réside dans le fait que l'ETE est avant tout centrée sur l'estimation de l'ampleur du travail des enfants au sein d'une *population de l'ensemble des enfants* définie de manière appropriée, et qu'en conséquence cette population a servi de base pour la conception et la sélection de l'échantillon d'ETE.

Pour le développement et l'exposé de la procédure d'échantillonnage de l'EET, nous supposons que le système d'échantillonnage de base est le suivant.

Nous partons de l'ETE. Un système d'échantillonnage très couramment utilisé pour l'ETE est celui du «plan avec PPT autopondérée» précédemment décrite: celle-ci comporte la sélection d'unités sectorielles avec probabilité proportionnelle à une mesure de la taille de population de la zone (p_i), puis, au sein de chaque zone sélectionnée, la sélection d'individus avec probabilité inversement proportionnelle à la mesure de la

taille. Nous nous référerons aux équations de sélection de base de ce plan précédemment obtenues:

$$f1_i = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}; \quad f2_i = \frac{b}{p_i}; \quad f = f1_i * f2_i = \frac{b}{I} \qquad [1]$$

Ici, *a* est le nombre des zones sélectionnées. Si p_i est strictement le nombre actuel des individus concernés dans chaque zone, *b* est le nombre constant d'unités finales sélectionnées dans chaque zone de sondage, et n=a*b la taille d'échantillon obtenue. $I = a/\Sigma p_i$ est l'intervalle qui serait utilisé dans une sélection systématique des zones avec PPT, la somme de valeurs p_i portant sur toutes les unités de la population. Chaque unité finale de la population a la même probabilité de sélection. Certaines variantes courantes de ce plan de base ont été décrites dans le chapitre 3[25]. Le plan de sondage peut renvoyer à une enquête sur la population générale telle que l'EMO, ou à une enquête sur le travail des enfants (ETE) de structure semblable à celle qui a été décrite dans le chapitre 4.

Les caractéristiques du plan de sondage

De la même manière, dans une enquête où la population de base concernée est constituée par les *enfants travailleurs*, un plan de sondage approprié comportera la sélection d'unités sectorielles avec probabilité proportionnelle au nombre d'enfants travailleurs (c_i) dans la zone, puis la sélection de ce type d'enfants à l'intérieur de chaque zone sélectionnée, avec probabilité inversement proportionnelle à c_i[26]:

$$f1'_i = \left(\frac{a'}{\Sigma c_i}\right).c_i = \frac{c_i}{I'}; \quad f2'_i = \frac{b'}{c_i}; \quad f' = f1'_i * f2'_i = \frac{b'}{I'} \qquad [2]$$

Ici, *a'* est le nombre de zones sélectionnées. Si c_i est strictement le nombre actuel des individus concernés (enfants travailleurs) dans chaque zone, *b'* est le nombre constant d'unités finales sélectionnées dans chaque zone de sondage, et n'=a'•b' la taille d'échantillon obtenue. La taille d'échantillon réelle à partir d'une grappe s'écartera de *b'* dans la mesure où c_i, mesure de la taille de la grappe, s'écarte de la taille réelle. $I' = \sum c_i / a'$ est l'intervalle qui serait utilisé dans la sélection systématique des zones avec PPT, la somme de valeurs c_i portant sur toutes les unités de la population. Dans cette conception de base, chaque enfant travailleur de la population a la même probabilité de sélection. Comme précédemment, certaines variations de cette conception de base sont possibles. Quelques-unes des considérations impliquées dans le choix de ces paramètres de conception ont été examinées dans les chapitres 2 et 3,

[25] Les procédures décrites dans ce chapitre sont facilement adaptables à d'autres plans, comme le «tirage fixe» par grappe, ou l'échantillonnage des secteurs avec probabilité constante plutôt qu'avec PPT. Les principes qui entrent en jeu ne présentent pas de différences.

[26] Ceci est un cas particulier de la structure plus générale illustrée dans le chapitre 6 (voir section 6.4.2), où la probabilité de sélection d'un secteur est considérée comme étant fonction de c_i comme de p_i. Nous avons choisi ici le modèle le plus simple pour mieux mettre en relief l'idée de base selon laquelle ce modèle revient à exclure les secteurs qui ne contiennent aucun enfant travailleur selon l'enquête de base (c_i=0).

mais nous supposerons pour le moment qu'elles ont été déterminées de manière appropriée.

La conception de l'EET renvoie à une enquête sur la population totale des enfants travailleurs, mais elle peut aussi être confinée aux enfants engagés dans des types spécifiques d'activités relevant du travail des enfants.

Cette conception diffère de celle de l'ETE examinée précédemment par un certain nombre d'aspects.

Premièrement, comme on l'a indiqué, la population de base et les mesures de taille sont c_i, le nombre des enfants engagés dans le travail des enfants. Ces valeurs ne peuvent pas être supposées connues pour toutes les zones de la population. On suppose qu'elles sont obtenues ou estimées dans la «première enquête» (EMO ou ETE par exemple), mais seulement pour les zones dénombrées dans cette enquête. Ensuite, l'enquête sur les enfants travailleurs doit être confinée à un sous-échantillon de zones de la première enquête. À l'intérieur de zones de sondage communes, les échantillons d'unités finales peuvent évidemment être différents ou coïncider.

Deuxièmement, la population de base des enfants travailleurs sera probablement répartie de manière beaucoup plus inégale dans les zones de sondage par comparaison avec la population générale des enfants. Quelques zones peuvent contenir des concentrations élevées, et beaucoup d'autres seulement des nombres très faibles d'enfants travailleurs. Il peut en particulier y avoir de nombreux «zéros», c'est-à-dire des zones ne contenant aucun enfant travailleur intéressant l'EET. Les problèmes tels que la présence de zones extrêmes («très grands» ou «très petits»), définis en fonction du nombre d'enfants travailleurs que contient la zone, seront probablement beaucoup plus répandus que ceux qui ont été évoqués précédemment pour l'ETE.

Troisièmement, la taille d'échantillon de l'EET sera probablement (ou du moins devrait être, dans une enquête de bonne qualité) beaucoup plus petite du fait de sa nature intensive.

5.2. Sélection des zones

Étant donné un échantillon de zones du type décrit par l'équation [1] dans la première enquête, comment le sous-échantillonner de manière à obtenir un échantillon de zones du type décrit par [2] pour l'EET ?

On peut y parvenir en sélectionnant un sous-échantillon de zones dans le premier échantillon avec PPT, l'indicateur de taille des zones pour ce sous-échantillonnage étant le ratio. Cela peut s'exprimer par:

$$g_i = a'^* \frac{(c_i / p_i)}{\Sigma_s (c_i / p_i)}, \quad f1_i^{'} = g_i \cdot f1_i \qquad [3]$$

où *a'* est le nombre de zones à sélectionner pour l'EET, la somme portant sur *toutes les zones du premier échantillon ou de l'échantillon de base (ETE ou EMO)* (comme

l'indique l'indice s). Ici g_i est la probabilité pour une zone (i) de l'échantillon de base d'être sélectionné pour l'EET; $f1'_i$ est la probabilité totale de sélection de la zone i pour l'EET, et $f1_i$ la même probabilité pour la zone de l'échantillon de base défini dans l'équation [1].

Cela donne, avec l'équation [1]:

$$f1'_i = \left[a * \frac{p_i}{\Sigma p_i}\right] * \left[a' * \frac{(c_i/p_i)}{\Sigma_s(c_i/p_i)}\right] = \left(\frac{a'}{k_s}\right) \cdot \frac{c_i}{\Sigma c_i}, \text{ où } k_s = \frac{\Sigma_s(c_i/p_i)/a}{\Sigma c_i/\Sigma p_i} \qquad [4]$$

Il s'ensuit que la procédure de sous-échantillonnage [3] aboutit à la sélection d'un échantillon de zones avec probabilité proportionnelle à la mesure de la taille c_i, selon les conditions requises.

Les quantités figurant dans l'équation ci-dessus sont les suivantes.

$f1'_i$ probabilité finale de la zone i dans l'échantillon d'EET. Elle est composée de deux facteurs figurant entre crochets. Le premier facteur est la probabilité de sélection de la zone dans l'échantillon de base. Le second est la probabilité de sélection de la zone depuis la base vers l'EET;

Σ somme portant sur les zones de la population;

Σ_s somme portant sur les zones de l'échantillon de base;

p_i mesure de la taille de la zone, telle qu'elle est utilisée pour la sélection de l'échantillon de base. Elle peut porter sur le nombre de ménages, de personnes ou d'enfants;

c_i mesure de la taille de la zone, telle qu'elle est utilisée pour la sélection de l'échantillon d'EET. Elle peut porter, par exemple, sur le nombre d'enfants travailleurs selon l'estimation de l'enquête de base, le nombre de ménages comportant des enfants travailleurs, ou toute autre mesure relative à l'extension du travail des enfants dans la zone. Les quantités c_i et p_i n'ont pas à être mesurées selon les mêmes unités ou la même échelle (puisque les équations en sont indépendantes sur le plan dimensionnel);

k_s constante déterminée par les caractéristiques de la population et de l'échantillon de base, indépendantes de l'échantillon d' EET ou de la zone particuliére i.

On notera que si l'échantillon d'EET de la grappe a' avait été sélectionné directement à partir de la population, avec une probabilité de sélection des zones proportionnelle à la mesure de la taille c_i (fonction du nombre d'enfants travailleurs dans la zone), l'équation de la sélection aurait été:

$$f1''_i = a' \cdot \left(c_i / \sum c_i\right)$$

Ainsi, k_s est un facteur qui annonce «indirectement» l'effet de la sélection de cet échantillon, via l'enquête de base. Si c_i est proportionnel à p_i dans toutes les zones, on constate que $k_s=1$.

L'équation [4] est en fait identique à [2], mis à part la présence du facteur k_s indiqué. Ce facteur n'est pas connu, puisque les valeurs de c_i ne sont pas connues pour toutes les zones de la population. Il dépend aussi de l'échantillon particulier qui se trouve être sélectionné dans la première enquête – par la suite [4] n'indique pas les «vraies» probabilités de sélection au sens de valeurs moyennes pour tous les échantillons de base possibles. Néanmoins, la valeur de ce facteur doit être proche de 1,0, puisque son numérateur comme son dénominateur sont des estimations du rapport c/p: le numérateur est la moyenne des ratios séparés c_i/p_i, cependant que le dénominateur est le ratio combiné des mêmes quantités. Dans tous les cas, ce facteur n'a pas d'incidence sur les probabilités *relatives* de sélection des unités sectorielles dans l'échantillon final, puisque le facteur est le même pour toutes ces unités. Elles sont donc sélectionnées avec une probabilité relative proportionnelle à leurs mesures de taille c_i.

En résumé, la procédure de sous-échantillonnage depuis des enquêtes existantes (EMO/ETE) plus importantes vers une EET plus restreinte s'effectue comme suit.

Des unités sectorielles de l'échantillon existant sont sous-échantillonnées avec la probabilité

$$\frac{(c_i/p_i)}{I_s}, \text{ où, } I_s = \frac{\Sigma_s (c_i/p_i)}{a'} \qquad [5]$$

Si bien que la probabilité de sélection réelle d'une zone dans le second échantillon est

$$f1'_i = f1_i . \frac{c_i/p_i}{I_s} = \left(\frac{p_i}{I}\right) \cdot \frac{c_i/p_i}{I_s} = \left(\frac{1}{I \cdot I_S}\right) \cdot c_i$$

Ce qui découle du fait que la probabilité de sélection de la zone concernée dans le premier échantillon est

$$f1_i = p_i/I.$$

5.3. Traiter les «très petites» et les «très grandes» zones

L'occurrence d'unités aux caractéristiques extrêmes sera probablement plus fréquente dans l'EET que dans des enquêtes comme l'EMO ou l'ETE. Il conviendra de veiller à ce que des probabilités de sélection correctes soient assurées pour ce type d'unités.

Bien entendu, ces problèmes relatifs à la taille des unités peuvent aussi se poser dans la sélection de l'échantillon de base. On supposera constamment qu'à ce stade, ces problèmes ont été traités comme nous l'avons expliqué dans les chapitres précédents: par exemple en redéfinissant p_i comme =I pour traiter les grandes unités, comme =b pour traiter les très petites unités, comme =0 pour les unités extrêmement petites (respectivement l'option [3] de la section 3.9 et l'option [3] de la section 3.10).

Les aspects indiqués ci-dessus concernant l'échantillon de base n'ont généralement pas pour conséquence de compliquer le traitement de ces cas dans l'EET. En effet, des ensembles de zones différents sont généralement impliqués en tant que cas «extrêmes» dans l'échantillon de base et celui de l'EET, puisque l'un et l'autre utilisent des types différents de mesures de taille, si bien qu'il est possible de traiter les deux ensembles séparément.

5.3.1. Les «très grandes» zones

Dans le contexte du sous-échantillonnage effectué à partir de l'échantillon de base pour obtenir un échantillon de zones pour l'EET, le terme «très grand» renvoie en fait non pas à la taille de la population, mais au *degré de concentration du travail des enfants*, c'est-à-dire aux valeurs très élevées du ratio (c_i/ p_i), puisque le ratio (c_i/p_i) est utilisé comme mesure de taille en [3]. Les très grandes unités sont celles pour lesquelles cette mesure excède l'intervalle d'échantillonnage utilisé dans la sélection des zones d'EET à partir de l'échantillon de base: $(c_i/p_i) \geq I_s$.

Ces zones peuvent être traitées de la même manière que dans les autres cas décrits dans les chapitres précédents. Ainsi, la mesure de taille (c_i/p_i)≥I est redéfinie comme $=I_s$, si bien que toutes ces zones du premier échantillon sont sélectionnées avec certitude dans l'échantillon d'EET.

Ces zones conservent donc des probabilités de sélection initiales dans l'échantillon de base inchangées pour l'EET. L'ensemble classé comme appartenant aux «grandes» zones dans l'EET est indépendant de l'ensemble classé comme tel dans l'enquête de base, puisque l'un et l'autre sont définis sur la base de critères de taille différents.

Quant au stade final, celui de la sélection de ménages ou de personnes dans les zones de sondage, le taux d'échantillonnage $f2'_i$ au stade final peut, par exemple, être ajusté de manière correspondante afin de maintenir inchangée la probabilité de sélection totale requise f'.

5.3.2. Les «très petites» zones

Pour les grandes zones, la procédure d'échantillonnage avec PPS ne nécessitait d'être ajustée que parce que la mesure de taille (c_i/p_i) excédait l'intervalle d'échantillonnage I_s. Les zones étaient donc identifiéees comme étant «grandes» ou non sur les bases du rapport (c_i/p_i).

Par contre, des zones sont définies comme étant «très petites» en fonction du nombre des unités finales qu'elles possèdent ou qu'elles apportent à l'échantillon, c'est-à-dire du nombre (ou du nombre moyen) des enfants travailleurs c_i.

La présence de petites valeurs de c_i a des conséquences pratiques importantes.

En premier lieu, il faut souligner que dans la conception décrite ci-dessus, celle de la sélection de zones de l'enquête de base avec probabilité proportionnelle à (c_i/p_i), les zones pour lesquelles aucun enfant travailleur n'a été signalé dans l'enquête de base (c_i=0) sont automatiquement exclues de l'échantillon d'EET. Sur un plan formel, il en va évidemment de même pour les zones sans population (p_i=0) quand on sélectionne

des zones de sondage dans l'échantillon de base avec probabilité proportionnelle à p_i. Cependant, dans la pratique, les deux situations sont très différentes. Les zones sans population (p_i=0) tendent à être rares et sans intérêt pour l'enquête en tout état de cause, alors que les zones sans enfants travailleurs (c_i=0) peuvent se rencontrer très couramment.

Deuxièmement, la situation concernant le deuxième cas (les enfants travailleurs) sera sans doute beaucoup plus changeante que celle du premier cas (la population). L'information sur la présence ou non d'enfants travailleurs dans une zone donnée doit donc être parfaitement actualisée.

Troisièmement, même quand elles ne sont pas exactement égales à zéro, les très petites valeurs de c_i ont une probabilité d'occurrence beaucoup plus importante que les petites valeurs de p_i, et cela à cause de la répartition inégale du travail des enfants entre les zones d' échantillon. La question pratique qui se pose est de savoir s'il faudrait imposer une limite inférieure pour l'exclusion automatique de l'échantillon des zones ayant des valeurs de c_i situées en dessous de cette limite.

Des procédures différentes peuvent être utilisées pour la sélection des ménages à l'intérieur des zones d'échantillon. Ces procédures sont probablement plus diverses dans les enquêtes sur les enfants travailleurs, par comparaison avec d'autres enquêtes plus générales sur la population.

Si les ménages sont sélectionnés à l'intérieur des zones d'échantillon avec PPT inverse, les procédures applicables aux très petites zones sont celles qui ont été décrites dans les chapitres précédents. Avec les «très petites» zones définies en fonction de la mesure de taille c_i, on considère comme «petites» les zones dont la mesure de taille est inférieure au tirage d'échantillon requis au stade final, autrement dit $c_i < b'$ dans l'équation relative à $f2'_i$ (équation [2]). Ces unités peuvent être traitées de la même manière que, par exemple, l'option (3) ou l'option (5) de la section 3.10.

Une préoccupation évidente de l'EET est d'assurer que la taille de l'échantillon d'enfants travailleurs requise soit réalisable dans la pratique. Cela peut poser un problème si l'ampleur du travail des enfants est moindre qu'on ne le supposait au moment de la conception de l'échantillon, ou si la médiocre qualité de l'enquête précédente n'a pas permis d'identifier une proportion élevée d'unités sujettes au travail des enfants. Quand ce type de problème existe, un plan de «grappe compacte» ou de «recensement» peut constituer une option intéressante. Dans ce plan, toutes les unités concernées (par ex. les ménages comportant des enfants travailleurs) dans une zone sélectionnée sont incluses dans l'échantillon, éventuellement avec une limite supérieure imposée au nombre maximum à sélectionner. Dans ce type de plan, les probabilités de sélection des unités finales, ainsi que les tirages d'échantillon par zone, sont généralement variables.

5.4. Augmenter la taille des premières zones d'échantillon

Il peut arriver que le type de zones initialement sélectionnées dans l'échantillon de base soit généralement trop petit pour offrir le nombre de cas requis pour l'EET. Dans ce type de situations, il peut être nécessaire de se demander si certaines des zones – peut-être celles qui comportent des concentrations élevées d'enfants travailleurs, et qui ont aussi probablement ces mêmes concentrations dans leur voisinage – peuvent être physiquement élargies pour inclure des zones voisines additionnelles. Nous décrirons ici une procédure simple de remplacement de zones d'échantillon existantes par des zones plus étendues.

Supposons que le type d'unités sectorielles utilisées pour la sélection du premier échantillon (A) soit composé de subdivisions d'unités de niveau supérieur (B), et que celles-ci soient considérées comme les unités appropriées pour l'échantillon élargi. Une zone (unité de niveau A) de l'échantillon existant peut être remplacée par l'unité plus importante de niveau B à laquelle elle appartient. L'échantillon résultant sera équivalent, statistiquement parlant, à la sélection d'unités entières de niveau B, *avec la probabilité de sélection d'une unité de niveau B tenue pour égale à la somme des probabilités de sélection de toutes les unités de niveau A qu'elle contient*. C'est avec cette probabilité accrue qu'apparaît, dans l'échantillon élargi, chaque unité de niveau A à l'intérieur de l'unité plus vaste de niveau B. Les unités plus vastes qui entrent dans l'échantillon de cette manière peuvent ensuite être dénombrées pour obtenir l'information requise (comme les mesures de taille c_i) pour la sélection d'un échantillon à partir de ces unités plus vastes en vue de l'EET[27].

Plus spécifiquement, la procédure est la suivante.

Supposons que la probabilité de sélection réelle d'une unité de niveau A dans le premier échantillon soit P_A/I, où P_A est sa mesure de taille, et I l'intervalle de sélection avec PPT. Chaque unité de niveau A de la population est supposée appartenir à une seule unité particulière de niveau B. La procédure implique de remplacer l'unité réelle de niveau A de l'échantillon par la totalité de l'unité de niveau B à laquelle elle appartient. La probabilité de sélection de l'unité de niveau B concernée est P_B/I, où P_B est la somme des mesures de taille de toutes les unités de niveau A contenues dans les unités de niveau B.

Par la suite, le sous-échantillonnage en vue de l'EET peut être appliqué à toute l'unité de niveau B «introduite» dans le premier échantillon de la manière exposée ci-dessus.

Il n'est pas nécessaire d'élargir ainsi toutes les unités, à condition que les critères déterminant les types d'unités à élargir et la procédure à suivre pour cette extension fassent l'objet d'une décision avant la sélection de l'échantillon et ne soient pas influencés par la nature des unités particulières de niveau A qui se trouvent sélectionnées dans l'échantillon.

[27] Si une unité de niveau B contient occasionnellement deux unités de niveau A sélectionnées dans l'échantillon original, elle peut alors être considérée comme ayant été sélectionnée deux fois dans l'échantillon élargi. Il en va de même pour une unité comptant plus de deux sélections.

Si la procédure simple décrite ci-dessus peut souvent être suffisante, des procédures plus complexes peuvent aussi être élaborées, avec pour objectif, par exemple, une extension des unités sectorielles conçue de telle sorte que la zone plus petite initialement sélectionnée se trouve au centre géographique de la zone élargie ou à proximité de celle-ci.

L'échantillonnage adapté par grappes est une autre approche plus élaborée qui peut être utile et réalisable dans certaines circonstances. Ses principes de base sont résumés dans la section suivante.

5.5. L'échantillonnage adapté par grappes

L'échantillonnage adapté

L'échantillonnage adapté est particulièrement utile quand la population concernée est rare, inégalement répartie, dissimulée ou difficile à atteindre. Dans l'échantillonnage conventionnel, le plan de sondage est entièrement basé sur l'information a priori et déterminé avant le début de l'étude. Au contraire, dans l'échantillonnage adapté, le plan de l'échantillonnage est adapté sur la base des observations faites en cours d'enquête[28].

L'objectif de cette procédure d'échantillonnage est de mettre à profit les caractéristiques de la population pour obtenir, par rapport aux conceptions conventionnelles, des estimations plus précises des valeurs de population pour une taille ou un coût d'échantillon donnés. L'objectif secondaire est d'accroître la production d'observations pertinentes qui favorisent une meilleure estimation d'autres paramètres les concernant. Cet aspect peut être extrêmement utile pour capter des échantillons assez larges de populations non nationales, par exemple.

Contrairement aux plans d'échantillonnage conventionnels, l'échantillonnage adapté fait usage des valeurs observées dans l'échantillon. Si l'échantillonnage séquentiel se préoccupe également des données, l'information ainsi obtenue est utilisée pour décider du nombre d'unités supplémentaires à échantillonner, ou bien de l'arrêt ou de la poursuite de l'échantillonnage. Par contre, l'échantillonnage adapté indique *quelles* unités il convient d'inclure dans l'échantillon.

Des procédures d'estimation particulières prenant en compte le plan d'échantillonnage sont nécessaires quand l'échantillonnage adapté a été utilisé. Ces procédures peuvent produire des estimations qui seront largement supérieures aux résultats conventionnels. Pour les populations rares et groupées, les modèles d'échantillonnage adapté peuvent permettre des gains d'efficience substantiels par rapport aux conceptions conventionnelles, et pour les populations cachées le suivi des liens et autres procédures adaptées peuvent constituer le seul moyen pratique d'obtenir un échantillon assez grand pour les objectifs de l'étude.

---

[28] Références de base:
Tompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association* 85, 1050-1059.
Tompson, S.K. (1992). *Sampling*. John Wiley & Sons, Inc., New York, 339 p.
Tompson, S.K., Seber. G.A.F. (1996*). Adaptive sampling*. John Wiley & Sons, Inc., New York, 265 p.

L'échantillonnage adapté par grappes

L'échantillonnage adapté représente en fait toute une catégorie d'approches particulières. Les modèles de suivi des liens tels que l'échantillonnage en boule de neige, les méthodes de marche aléatoire et l'échantillonnage en réseau, la répartition adaptée et l'échantillonnage *adapté par grappes* représentent des formes diverses de plans d'échantillonnage adapté.

Le type d'échantillonnage adapté le plus souvent mentionné est appelé *échantillonnage adapté par grappes*. Celui-ci est utile dans les situations où la caractéristique intéressante est répartie de manière clairsemée mais fortement concentrée. Dans l'échantillonnage environnemental, par exemple, on peut trouver des exemples de ces populations dans l'étude d'espèces rares et menacées, ou de concentrations de pollution. Dans les populations humaines, ces techniques peuvent être utiles dans les recherches sur les pires formes de travail des enfants, l'épidémiologie des maladies rares, les concentrations de populations d'immigrants, etc.

L'échantillonnage adapté par grappes est extrêmement utile quand une rotation rapide des résultats analytiques est possible, puisque la poursuite de l'échantillonnage dépend à tout moment de l'analyse de l'information déjà collectée.

L'échantillonnage adapté par grappes implique que soient spécifiées:

1. le plan et la taille de l'échantillonnage initial (avant tout apport de l'échantillonnage adapté)
2. la définition de ce qui constitue le «voisinage» pour une unité d'échantillonnage
3. la condition qui déclenche ou impulse l'échantillonnage adapté pour une unité de l'échantillon initial
4. toute restriction à l'application répétée des règles concernant l'ajout de nouvelles unités dans l'échantillon (afin de contrôler la taille d'échantillon obtenue)
5. les procédures d'estimation (y compris l'estimation de variance).

Nous commencerons par un échantillon initial 1, sélectionné selon les procédures conventionnelles, et les unités constituant le voisinage de chaque unité sélectionnée selon la condition 2. Si les caractéristiques d'une unité échantillonnée satisfont à la condition 3 de l'échantillonnage adapté, toutes les unités du voisinage de l'unité déjà sélectionnée sont ajoutées à l'échantillon. Si l'une quelconque des unités nouvellement ajoutées dans ce voisinage satisfait à la condition 3, le voisinage de cette unité défini selon 2 est également ajouté à l'échantillon. Le processus se poursuit jusqu'à obtention d'une grappe d'unités délimitée par des unités «limitrophes» qui ne satisfont pas à la condition 2, ou qui sont sujettes à restriction 4. L'échantillon final se compose de grappes (qui ne sont pas nécessairement distinctes), une pour chaque unité sélectionnée dans l'échantillon initial. Les procédures d'estimation 5 prennent en compte les probabilités d'apparition dans l'échantillon de l'unité initiale et de celle qui a été ajoutée par la suite, ainsi que la structure de l'échantillon résultant.

Les points forts

1. Premièrement, à la différence des plans traditionnels qui se concentrent sur un seul objectif, l'échantillonnage adapté vise à atteindre simultanément l'objectif consistant à estimer la concentration moyenne et à déterminer le modèle et l'ampleur des concentrations du phénomène étudié.

2. L'échantillonnage *adapté par grappes* concentre les ressources sur des zones et des types d'unités/événements d'un intérêt particulier: sur les zones à plus forte concentration de ces unités ou de ces événements. Il dirige la sélection d'unités d'échantillonnage additionnelles vers ces zones à forte concentration, sous réserve que l'échantillon initial «touche» les zones intéressantes.

3. De plus, on peut observer des caractéristiques additionnelles qui ajoutent à l'utilité globale de l'étude. Par exemple, dans les études portant sur la présence ou l'absence de populations animales rares, des mesures de taille, de poids, etc., peuvent être effectuées sur les animaux que l'on découvre. Il en va de même pour les enquêtes sociales sur des populations humaines, par exemple sur *les pires formes de travail des enfants*.

Les limites

1. La nature itérative de l'échantillonnage *adapté par grappes* introduit certaines limites. Avec cette méthode, le processus consistant à échantillonner, tester, rééchantillonner et tester à nouveau peut demander un temps considérable. Si des mesures rapides et peu coûteuses ne sont pas aisément disponibles sur le terrain, le coût total de l'échantillonnage risque de devenir très vite important.

2. Du fait que le processus d'échantillonnage s'arrête seulement lorsqu'on ne trouve plus d'unités présentant les caractéristiques qui intéressent l'enquête (ou qu'une autre restriction spécifiée devient applicable), la taille d'échantillon finale est une quantité inconnue. Cet aspect fait également du coût total une quantité inconnue.

3. Bien qu'il soit possible de budgétiser le processus d'échantillonnage en utilisant le coût total attendu, ce dernier dépend aussi dans une large mesure de la validité de l'hypothèse formulée sur l'extension de la caractéristique étudiée. Si celle-ci est très largement répandue, la taille d'échantillon obtenue peut «exploser» au point de devenir ingérable. Considérons une enquête sur une population rare pour laquelle on suppose qu'il existe seulement quelques petites zones de concentration élevée. Supposons que cette hypothèse ne soit pas fondée, autrement dit, que la concentration soit plus diffuse et concerne la quasi-totalité de la zone étudiée. L'échantillon initial a une forte probabilité de «toucher» une zone de concentration. Les zones de concentration étant largement étendues, la taille de l'échantillon complémentaire sera importante: elle pourra même approcher le nombre d'unités d'échantillonnage de la population totale.

4. D'un autre côté, la taille d'échantillon peut se révéler inadéquate au but de l'enquête si le phénomène étudié est plus concentré que prévu et que l'échantillon initial ne soit pas assez important pour le capter de manière adéquate.

5. Généralement, les procédures d'estimation sont plus (souvent beaucoup plus) complexes que celles qui concernent un plan conventionnel.

Exemple

Considérons le scénario suivant pour l'étude d'une sous-population qui est très inégalement répartie dans la population générale. Supposons que dans la plupart des endroits échantillonnés, la concentration de cette sous-population soit faible ou négligeable, mais que l'on rencontre quelques poches dispersées de forte concentration. Deux questions sont à étudier. Premièrement, quel est le niveau moyen de concentration pour l'ensemble de la zone – autrement dit, la proportion appartenant à cette sous-population ? Deuxièmement, où les poches de concentration sont-elles localisées et quelles sont leurs caractéristiques ?

Si l'on utilisait la méthode statistique traditionnelle, on sélectionnerait un échantillon aléatoire ou systématique des sites et des unités, et l'on mesurerait la concentration sur chaque site sélectionné. La moyenne de ces mesures fournit une estimation non biaisée de la moyenne de population. Les observations individuelles peuvent être utilisées pour créer des courbes de niveau afin de localiser les pics de concentration. Toutefois, avec ce modèle de concentration, l'approche statistique traditionnelle peut rencontrer des problèmes. Si la concentration est négligeable sur la plus grande partie de la zone, la majorité des mesures seront égales à zéro ou se situeront à des niveaux non décelables. De plus, l'échantillonnage aléatoire peut «manquer» la plupart des poches de concentration plus élevée.

Quand bien même l'échantillon moyen serait un estimateur sans biais de la moyenne de population, il sera moins précis qu'un estimateur sans biais prenant en compte le caractère inégal de la répartition de la sous-population sur la zone tout entière. En outre, les courbes de niveau établies à partir d'un plan d'échantillon aléatoire simple peuvent être moins exactes dans les zones de forte concentration, insuffisamment représentées dans l'échantillon.

L'échantillonnage adapté par grappes pourrait offrir une meilleure approche dans des situations semblables à celle qui a été décrite ci-dessus. Pour les populations dans lesquelles la caractéristique étudiée est répartie de manière clairsemée mais fortement concentrée, il peut apporter des gains substantiels en précision par rapport aux plans d'échantillonnage traditionnels utilisant les mêmes tailles d'échantillon.

L'échantillonnage adapté est particulièrement utile quand la population étudiée est rare, inégalement répartie, cachée ou difficile à atteindre[29]. Parmi les exemples de ces populations figurent les utilisateurs de drogues par injection, les personnes exposées à un risque élevé de contamination par le VIH/SIDA, et les jeunes adolescents dépendants de la nicotine. Dans l'échantillonnage conventionnel, le plan d'échantillonnage est entièrement basé sur l'information a priori, et elle est déterminée avant le début de l'étude. Par contre, dans l'échantillonnage adapté, le plan de l'échantillonnage s'adapte sur la base d'observations effectuées au cours de l'enquête; par exemple, on peut demander aux utilisateurs de drogues de renvoyer d'autres utilisateurs de drogues aux enquêteurs. Des procédures d'estimation spéciales prenant en compte la conception de l'échantillonnage sont nécessaires quand l'échantillonnage adapté a été utilisé.

[29] Thompson S.K., Collins L.M. (2002). Adaptive sampling in research on risk-related behaviours. *Drug and alcohol dependence* 68 (2002) S57-S67. Cet article présente des plans d'échantillonnage adapté à des chercheurs sur l'utilisation de substances. Le texte est le résumé de l'article.

5.6. Illustrations numériques

L'objectif de ces simulations est d'illustrer par des chiffres certains aspects des procédures de plan et de sélection d'échantillons examinées dans ce chapitre.

Les illustrations chiffrées sont basées sur une population d'unités sectorielles présentée dans la section 3.11.1. Comme nous l'avons indiqué, un petit ensemble de données a été produit statistiquement pour fournir un exemple raisonnablement réaliste d'un ensemble de zones avec leurs mesures de taille associées. La population initiale a été constituée de quelques P=50.000 unités finales (ménages), dans près de 800 zones d'une taille moyenne d'environ 65 ménages. Les zones variaient considérablement en taille, depuis le plus petit, comptant moins de 10 unités, jusqu'au plus grand, avec plus de 1 000 unités. La répartition était assez typique des enquêtes réelles, fortement asymétrique vers la gauche avec beaucoup de petites zones et quelques très grandes zones.

Un échantillon systématique avec PPT de a=100 zones a été sélectionné, l'intervalle d'échantillonnage pour la sélection des zones étant I=(P/a)=500.

Pour chaque zone de l'ensemble a également été simulé c_i, le nombre d'enfants travailleurs qu'il contenait. Dans une enquête réelle, ce type d'information serait obtenu dans une ETE précédant l'EET. Dans le modèle examiné ici, les valeurs de c_i (ou plus précisément leur ratio c_i/p_i par rapport à la population de la zone) déterminent les probabilités de sélection des zones de l'ETE vers l'EET.

Les colonnes [6]-[7] du tableau 3.3 indiquaient les valeurs simulées de c_i, le nombre d'enfants travailleurs dans la zone *i*, et le ratio de celui-ci par rapport à p_i, la population d'enfants du groupe d'âge concerné exposés au risque de travail des enfants. Ces valeurs illustraient la large amplitude du degré de concentration du travail des enfants selon les zones: la proportion des enfants engagés dans le travail des enfants variait de 0 pour cent à 97 pour cent, la moyenne d'ensemble étant de 22 pour cent.

5.6.1. La sélection de zones à partir de l'échantillon de base

L'échantillon de 100 zones présenté dans le tableau 3.3 a été ordonné selon les valeurs de p_i (mesure simulée de la population ou du nombre de ménages). On a supposé que les zones de sondage avaient été sélectionnées avec probabilité proportionnelle à cette mesure de taille. Les zones ont été numérotées séquentiellement de 1 à 100 dans l'ordre mentionné ci-dessus. Les mêmes 100 zones sont classées selon des ordres différents dans les tableaux 5.1-5.3. Le numéro de série de la première colonne de chacun de ces tableaux identifie la numérotation originale du tableau 3.3, si bien que les informations relatives à n'importe quelle zone donnée peuvent être suivies d'un tableau à l'autre.

Dans la sélection initiale des zones, un traitement spécial a été prévu pour les zones classées comme étant «trop grandes» ou «trop petites», comme nous l'avons expliqué dans les notes du tableau 3.3 (section 3.11). Les zones dont les mesures de taille excédaient l'intervalle de l'échantillonnage systématique (I=500) ont reçu une mesure de taille modifiée de 500 et une probabilité de sélection de 1,0. Les zones dont les

mesures de taille étaient plus petites que le tirage d'échantillon cible supposé (b=20 ménages) ont reçu une mesure de taille modifiée de 20, et leur probabilité de sélection a été considérée comme proportionnelle à cette mesure de taille modifiée.

Ces ajustements ont été maintenus aux stades suivants de la sélection de l'ETE à l'EET illustrée dans les tableaux 5.1-5.3. Ils n'ont pas d'autres incidences sur les procédures appliquées dans ces tableaux.

Le cumul de (c_i/p_i) sur les 100 zones (25,8), divisé par le nombre des zones à sélectionner (50), donne l'intervalle à appliquer pour une sélection systématique des zones, I=0,52.

Les zones avec (c_i/p_i)>I sont maintenues dans l'échantillon d'EET avec une probabilité de 1,0 (au total, 18 zones sont apparues «très grandes» dans ce sens – voir tableau 5.1). Pour ces zones c_i est redéfini comme $c'_i=I \bullet p_i$ (tableau 5.1, col. [7]), c'est-à-dire comme (c'_i/p_i)=I, afin de garantir que la probabilité de sélection n'excède pas 1,0. Dans toutes les autres zones, il n'y a pas de changement: $c_i= c_i$ initial.

Les probabilités de sélection des zones finales à partir de la base pour l'échantillon d'EET (f1) sont indiquées dans la col. [6]; elles sont proportionnelles à (c'_i/p_i). La probabilité de sélection globale d'une zone est $f0_i \bullet f1_i$, le premier chiffre provenant de la col. (5) du tableau 3.3.

On notera qu'il ne s'agit pas là du stade final de la sélection de zones pour l'hypothétique EET qui est envisagée. Dans le tableau, les valeurs de (c'_i/p_i) ou de c_i indiquent un problème pratique courant: de nombreuses zones ne comportent pas, ou très peu, d'enfants travailleurs, et la question est de savoir s'il l'on peut se permettre d'intégrer ces zones dans l'échantillon. Ces «très petites zones» ont besoin d'un traitement particulier[30], dont les détails dépendent de la procédure adoptée pour le second stade de l'échantillonnage: la sélection des unités finales à l'intérieur des zones d'échantillonage. On en trouvera l'illustration dans le tableau 5.2.

[30] On notera que l'adjectif «petit» ne s'applique pas ici à la taille de population (p_i) du secteur, mais à la variable qui nous intéresse dans l'EET, à savoir le nombre d'enfants travailleurs (c_i).

Tableau 5.1: La sélection de zones à partir de l'échantillon de base

Paramétres:

Enquête de base , a = 100 [nombre de zones dans le base de sondage]
EET, a = 50 [nombre de zones dans l'échantillon EET]
I = 1,9 [intervalle pour la sélection systématique des zones]
n = 500 [taille d'échantillon espérée]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	(modifié)		(classé selon)		(modifié)	(modifié)
S.No	pi	ci	ci/pi	(ci/pi)/I	f1	ci
73	196	190	0,97	1,88	1,00	101
54	100	93	0,92	1,79	1,00	52
92	425	382	0,90	1,74	1,00	219
84	299	268	0,90	1,74	1,00	154
49	83	74	0,90	1,74	1,00	43
35	60	53	0,89	1,73	1,00	31
42	70	61	0,87	1,69	1,00	36
10	22	18	0,81	1,57	1,00	11
97	500	383	0,77	1,48	1,00	258
65	157	117	0,74	1,44	1,00	81
55	103	73	0,71	1,38	1,00	53
45	73	51	0,69	1,34	1,00	38
86	331	220	0,66	1,29	1,00	171
7	20	12	0,62	1,20	1,00	10
25	43	27	0,62	1,20	1,00	22
60	115	68	0,60	1,15	1,00	59
93	500	292	0,58	1,13	1,00	258
47	79	41	0,52	1,01	1,00	41
26	44	23	0,51	0,99	0,99	23
28	48	23	0,48	0,94	0,94	23
44	72	35	0,48	0,93	0,93	35
78	229	107	0,47	0,90	0,90	107
24	43	20	0,46	0,90	0,90	20
12	25	11	0,45	0,87	0,87	11
20	38	17	0,45	0,86	0,86	17
89	368	162	0,44	0,85	0,85	162
3	20	9	0,44	0,85	0,85	9
1	20	9	0,44	0,85	0,85	9
27	46	18	0,39	0,75	0,75	18
67	163	51	0,31	0,61	0,61	51
18	35	11	0,31	0,61	0,61	11
46	74	20	0,27	0,52	0,52	20
2	20	5	0,26	0,51	0,51	5
76	213	55	0,26	0,50	0,50	55
8	20	5	0,24	0,47	0,47	5
52	97	22	0,23	0,44	0,44	22
70	181	40	0,22	0,42	0,42	40
58	114	25	0,22	0,42	0,42	25
61	124	27	0,21	0,42	0,42	27

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	(modifié)		(classé selon)		(modifié)	(modifié)
S.No	pi	ci	ci/pi	(ci/pi)/I	f1	ci
51	92	20	0,21	0,42	0,42	20
21	39	8	0,21	0,41	0,41	8
15	30	6	0,21	0,40	0,40	6
32	53	11	0,20	0,40	0,40	11
57	112	21	0,19	0,37	0,37	21
94	500	92	0,18	0,35	0,35	92
79	229	42	0,18	0,35	0,35	42
37	65	12	0,18	0,35	0,35	12
59	114	20	0,18	0,35	0,35	20
36	64	11	0,17	0,34	0,34	11
38	67	11	0,17	0,33	0,33	11
33	57	9	0,16	0,31	0,31	9
53	97	15	0,16	0,31	0,31	15
16	33	5	0,15	0,30	0,30	5
56	107	16	0,15	0,29	0,29	16
29	50	7	0,15	0,28	0,28	7
5	20	3	0,14	0,27	0,27	3
68	171	22	0,13	0,25	0,25	22
71	186	23	0,12	0,24	0,24	23
96	500	60	0,12	0,23	0,23	60
85	304	36	0,12	0,23	0,23	36
19	36	4	0,11	0,21	0,21	4
72	189	20	0,11	0,21	0,21	20
64	156	16	0,11	0,20	0,20	16
48	79	7	0,09	0,18	0,18	7
100	500	39	0,08	0,15	0,15	39
88	358	26	0,07	0,14	0,14	26
69	175	11	0,06	0,12	0,12	11
43	72	3	0,04	0,08	0,08	3
40	69	3	0,04	0,08	0,08	3
75	204	8	0,04	0,07	0,07	8
80	237	9	0,04	0,07	0,07	9
39	68	2	0,03	0,07	0,07	2
82	240	7	0,03	0,06	0,06	7
6	20	1	0,03	0,05	0,05	1
63	138	3	0,03	0,05	0,05	3
13	28	1	0,03	0,05	0,05	1
99	500	9	0,02	0,03	0,03	9
41	69	1	0,02	0,03	0,03	1
14	28	0	0,02	0,03	0,03	0
30	52	1	0,01	0,03	0,03	1
22	40	1	0,01	0,03	0,03	1
11	24	0	0,01	0,02	0,02	0
50	92	1	0,01	0,02	0,02	1
81	238	2	0,01	0,02	0,02	2

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	(modifié)		(classé selon)		(modifié)	(modifié)
S.No	pi	ci	ci/pi	(ci/pi)/I	f1	ci
17	34	0	0,01	0,02	0,02	0
66	159	1	0,01	0,01	0,01	1
90	383	2	0,01	0,01	0,01	2
77	213	1	0,01	0,01	0,01	1
31	52	0	0,01	0,01	0,01	0
87	345	1	0,00	0,01	0,01	1
23	40	0	0,00	0,01	0,01	0
83	246	1	0,00	0,01	0,01	1
74	200	1	0,00	0,00	0,00	1
34	59	0	0,00	0,00	0,00	0
4	20	0	0,00	0,00	0,00	0
95	500	0	0,00	0,00	0,00	0
98	500	0	0,00	0,00	0,00	0
91	424	0	0,00	0,00	0,00	0
62	128	0	0,00	0,00	0,00	0
9	21	0	0,00	0,00	0,00	0
Moyenne	150,0	37,5	25,79	50,00		29,6

5.6.2. La sélection d'unités finales et le traitement des «très petites» zones

Afin d'illustrer la procédure suivie, supposons que les unités finales soient sélectionnées à l'intérieur des zones d'échantillonage avec une PPT inverse afin d'obtenir un échantillon autopondéré[31]. D'autres modèles sont évidemment possibles. Avec le plan de sondage avec PPT, supposons qu'une moyenne de b=10 enfants travailleurs doit être sélectionnée par zone de sondage, ce qui donnera une taille d'échantillon totale espérée de n=a*b=50*10=500 enfants travailleurs.

La col. [5] indique les valeurs de $c_{i;}$ ce sont les mêmes que dans la col. [7] du tableau 5.1, si ce n'est que les données sont maintenant classées par ordre croissant des valeurs de c_i, puisque notre objectif est d'identifier les zones qui sont «très petites» en regard de cette mesure. La col. [5] identifie les zones qui sont «très petites» en ce sens qu'elles contiennent moins de b=10 enfants travailleurs. Dans la population qui illustre notre propos, les zones de ce type sont nombreuses (45 sur 100). Cela n'est pas irréaliste, car les enfants travailleurs sont souvent concentrés géographiquement, et ils sont inexistants ou très rares dans de nombreuses zones.

Diverses procédures relatives à la sélection des «très petites» zones ont été décrites dans la section 3.10. La méthode (5) figurant dans cette section apparaît comme satisfaisante et a été appliquée dans cette illustration.

[31] Il est sans incidence, pour le sujet exposé ici, que les unités finales puissent être définies de différentes manières – il peut s'agir par exemple des ménages comportant un ou plusieurs enfants travailleurs, ou des enfants travailleurs considérés individuellement ; on pourra inclure, pour faciliter les comparaisons, quelques enfants non travailleurs ou des ménages sans aucun enfant travailleur.

Cette méthode consiste à trier l'ensemble de très petites zones par mesure de taille (c_i), et à le diviser en deux parties. La première partie est composée des plus grands parmi cet ensemble de très petites zones. Le nombre de zones (disons c) à inclure dans cette partie est déterminé de telle manière que si chacune de ces unités recevait une mesure de taille b (=10, l'échantillon cible tiré par zone), cela représenterait la mesure de taille totale (disons $?_s c_i$) de toutes les unités du «très petit ensemble». Autrement dit,

$c=\Sigma_s c_i/b=138/10=14$ dans notre illustration.

En vue de la sélection de l'échantillon, on assigne à chacune de ces 14 unités une mesure de taille de 10. Les zones restantes (45-14=31 zones) du «très petit» ensemble sont réellement très petites, et elles sont entièrement exclues de l'échantillonnage par l'attribution d'une mesure de taille égale à zéro. Les mesures de taille ajustées sont indiquées dans la col. [6].

La pondération finale pour les 14 unités maintenues, si elles sont sélectionnées, peut être ajustée vers le haut en compensation des 31 exclues parmi les plus petites unités. On en trouvera l'illustration dans le tableau 5.3.

Le tableau 5.2 présente également les probabilités de sélection à des stades divers:

$f0_i$ pour la sélection initiale des zones d'ETE (provenant du tableau 3.3)

$f1_i$ pour la sélection des zones de l'ETE vers l'EET (provenant du tableau 5.1)

$f2_i$ pour la sélection des unités finales dans une zone $=10/c_i$, avec la mesure de taille c_i modifiée comme nous l'avons expliqué ci-dessus.

Tableau 5.2: La sélection d'unités finales dans les zones sélectionnées

Paramétres:

Enquête de base , a = 100 [nombre de zones dans le base de sondage]
EET, a = 50 [nombre de zones dans l'échantillon EET]
I = 1,9 [intervalle pour la sélection systématique des zones]
n = 500 [taille d'échantillon espérée]

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Échantillon de base		Sélection de zone	Classé par	Modifié-2	Sélection des ménages
Sno	pi	f0	f1	ci (modifié-1)	ci	f2
9	21	0,04	0,00	0	0	
62	128	0,26	0,00	0	0	
91	424	0,85	0,00	0	0	
4	20	0,04	0,00	0	0	
98	500	1,00	0,00	0	0	
34	59	0,12	0,00	0	0	
23	40	0,08	0,01	0	0	
95	500	1,00	0,00	0	0	
31	52	0,10	0,01	0	0	
11	24	0,05	0,02	0	0	
17	34	0,07	0,02	0	0	
14	28	0,06	0,03	0	0	

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Échantillon de base		Sélection de zone	Classé par	Modifié-2	Sélection des ménages
Sno	pi	f0	f1	ci (modifié-1)	ci	f2
74	200	0,40	0,00	1	0	
6	20	0,04	0,05	1	0	
22	40	0,08	0,03	1	0	
13	28	0,06	0,05	1	0	
83	246	0,49	0,01	1	0	
30	52	0,10	0,03	1	0	
66	159	0,32	0,01	1	0	
50	92	0,18	0,02	1	0	
87	345	0,69	0,01	1	0	
41	69	0,14	0,03	1	0	
77	213	0,43	0,01	1	0	
39	68	0,14	0,07	2	0	
90	383	0,77	0,01	2	0	
81	238	0,48	0,02	2	0	
40	69	0,14	0,08	3	0	
5	20	0,04	0,27	3	0	
43	72	0,14	0,08	3	0	
63	138	0,28	0,05	3	0	
19	36	0,07	0,21	4	0	
8	20	0,04	0,47	5	10	1,00
16	33	0,07	0,30	5	10	1,00
2	20	0,04	0,51	5	10	1,00
15	30	0,06	0,40	6	10	1,00
48	79	0,16	0,18	7	10	1,00
29	50	0,10	0,28	7	10	1,00
82	240	0,48	0,06	7	10	1,00
75	204	0,41	0,07	8	10	1,00
21	39	0,08	0,41	8	10	1,00
80	237	0,47	0,07	9	10	1,00
1	20	0,04	0,85	9	10	1,00
3	20	0,04	0,85	9	10	1,00
99	500	1,00	0,03	9	10	1,00
33	57	0,11	0,31	9	10	1,00
7	20	0,04	1,00	10	10	0,97
69	175	0,35	0,12	11	11	0,94
32	53	0,11	0,40	11	11	0,92
18	35	0,07	0,61	11	11	0,92
36	64	0,13	0,34	11	11	0,91
38	67	0,13	0,33	11	11	0,89
10	22	0,04	1,00	11	11	0,89
12	25	0,05	0,87	11	11	0,88
37	65	0,13	0,35	12	12	0,85
53	97	0,19	0,31	15	15	0,65
56	107	0,21	0,29	16	16	0,63

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Échantillon de base		Sélection de zone	Classé par	Modifié-2	Sélection des ménages
Sno	pi	f0	f1	ci (modifié-1)	ci	f2
64	156	0,31	0,20	16	16	0,61
20	38	0,08	0,86	17	17	0,60
27	46	0,09	0,75	18	18	0,56
51	92	0,18	0,42	20	20	0,51
46	74	0,15	0,52	20	20	0,51
24	43	0,09	0,90	20	20	0,51
72	189	0,38	0,21	20	20	0,50
59	114	0,23	0,35	20	20	0,49
57	112	0,22	0,37	21	21	0,47
52	97	0,19	0,44	22	22	0,46
68	171	0,34	0,25	22	22	0,45
25	43	0,09	1,00	22	22	0,45
26	44	0,09	0,99	23	23	0,44
71	186	0,37	0,24	23	23	0,43
28	48	0,10	0,94	23	23	0,43
58	114	0,23	0,42	25	25	0,41
88	358	0,72	0,14	26	26	0,39
61	124	0,25	0,42	27	27	0,38
35	60	0,12	1,00	31	31	0,32
44	72	0,14	0,93	35	35	0,29
42	70	0,14	1,00	36	36	0,28
85	304	0,61	0,23	36	36	0,28
45	73	0,15	1,00	38	38	0,26
100	500	1,00	0,15	39	39	0,26
70	181	0,36	0,42	40	40	0,25
47	79	0,16	1,00	41	41	0,25
79	229	0,46	0,35	42	42	0,24
49	83	0,17	1,00	43	43	0,23
67	163	0,33	0,61	51	51	0,20
54	100	0,20	1,00	52	52	0,19
55	103	0,21	1,00	53	53	0,19
76	213	0,43	0,50	55	55	0,18
60	115	0,23	1,00	59	59	0,17
96	500	1,00	0,23	60	60	0,17
65	157	0,31	1,00	81	81	0,12
94	500	1,00	0,35	92	92	0,11
73	196	0,39	1,00	101	101	0,10
78	229	0,46	0,90	107	107	0,09
84	299	0,60	1,00	154	154	0,06
89	368	0,74	0,85	162	162	0,06
86	331	0,66	1,00	171	171	0,06
92	425	0,85	1,00	219	219	0,05
93	500	1,00	1,00	258	258	0,04
97	500	1,00	1,00	258	258	0,04

[1]	[2]	[3]	[4]	[5]	[6]	[7]
	Échantillon de base		Sélection de zone	Classé par	Modifié-2	Sélection des ménages
Sno	pi	f0	f1	ci (modifié-1)	ci	f2
Moyenne	150,0			29,6	29,7	

5.6.3. Les probabilités de sélection globales et la taille d'échantillon attendue

Le tableau 5.3 présente les résultats finals. Il est classé, comme le tableau précédent, en fonction de c_i. L'illustration ne sélectionne pas réellement un échantillon particulier de a=50 zones, mais elle indique les probabilités de sélection de chacun des 100 zones, et l'échantillon qu'apporterait une zone s'il était sélectionné.

La col. [4] est celle du taux d'échantillonnage global f, soit le produit des probabilités de sélection à tous les stades indiqués dans les col. [1]-[3]: sélection dans les zones d'échantillon de base (f0); à partir de ceux-ci, sélection des zones de l'EET (f1); et à l'intérieur de ceux-ci, sélection des unités finales, par exemple des enfants travailleurs (f2).

La col. [5] indique les pondérations des 14 plus grandes zones retenues pour la sélection de l'échantillon parmi le «très petit» ensemble de 45 unités examiné ci-dessus en relation avec le tableau 5.2. Ces pondérations sont supérieures à 1,0 pour compenser l'exclusion des 31 plus petites zones restantes dans le «très petit» ensemble. Elles égalent en fait $(b/c_i)=(10/c_i)$, c_i étant la mesure de taille de l'unité, par définition pour toute unité du «très petit» ensemble. Les 31 zones exclues reçoivent un poids spécial égal à zéro. Par définition, le poids spécial est de 1,0 pour toutes les zones restantes qui ne sont pas dans le «très petit» ensemble.

La col. [6] indique le tirage d'échantillon (b_i) qui serait obtenu si l'unité était sélectionnée dans l'échantillon, en supposant que le nombre des unités finales (disons c'_i) réellement constaté dans la zone au moment de l'enquête ait été exactement le même que le nombre (c_i) utilisé dans la sélection de l'échantillon.

Pour la plupart des zones, le nombre de la col. [6] égale 10, car l'équation de sélection du stade final est $f2_i=10/c_i$, ce qui donne le nombre attendu de sélections, soit $f2_i \bullet c_i=10$.

Ce nombre diffère de 10 pour deux groupes d'unités:

- Les «très grandes» zones identifiées dans le tableau 5.1, dont la mesure de taille (c_i/p_i) a excédé l'intervalle d'échantillonnage des zones (I=0,52). Le ratio de ces deux quantités est le facteur par lequel leur contribution finale excède 10.
- Toute zone «très petite» qui a été retenue pour la sélection de l'échantillon a, par définition, moins de 10 unités à apporter. En fait, chaque zone de ce type apporte une contribution de (10/D) unités finales, D étant la pondération indiquée dans la col. [5]. On notera que cela représente une contribution *pondérée* de (10/D)•D=10 dans le cas de toutes les unités de ce type.

Enfin, la col. [7] présente l'espérance mathématique de l'échantillon apportée par chaque zone. Il s'agit de la contribution qu'apporterait en moyenne la zone, sur tous les échantillons possibles. Elle est plus petite que la contribution réelle qu'apporterait une zone une fois réellement apparue dans l'échantillon (col. [6]), mais on rappellera qu'il existe aussi de nombreuses occasions où la zone n'apporte aucune contribution réelle, parce qu'elle n'est pas sélectionnée dans un échantillon. Les valeurs attendues indiquées dans la col. [7] nous permettent d'examiner les caractéristiques moyennes des échantillons qui sont possibles avec un plan de sondage donné, sans devoir réellement sélectionner des échantillons individuels.

La somme portant sur l'ensemble des 100 zones de la «population» de valeurs de la col. [7] indique, par exemple, que le nombre attendu d'unités finales de l'échantillon est de 500.

Tableau 5.3: Les probabilités de sélection globales et la taille d'échantillon attendue

Paramétres:

Enquête de base , a = 100 [nombre de zones dans le base de sondage]
EET, a = 50 [nombre de zones dans l'échantillon EET]
I = 1,9 [intervalle pour la sélection systématique des zones]
n = 500 [taille d'échantillon espérée]

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Classé selon	Probabilité de sélection par niveau					Numéro des unités finales sélectionnées	
	Original				Final	Special	Si dans l'échantillon	Valeur espérée
Sno	ci	f0*	f1*	f2	=f	Pondérations (D)	bi	bi*D*f1
9	0	0,04	0,00					0
62	0	0,26	0,00					0
91	0	0,85	0,00					0
4	0	0,04	0,00					0
98	0	1,00	0,00					0
34	0	0,12	0,00					0
23	0	0,08	0,01					0
95	0	1,00	0,00					0
31	0	0,10	0,01					0
11	0	0,05	0,02					0
17	0	0,07	0,02					0
14	0	0,06	0,03					0
74	1	0,40	0,00					0
6	1	0,04	0,05					0
22	1	0,08	0,03					0
13	1	0,06	0,05					0
83	1	0,49	0,01					0
30	1	0,10	0,03					0
66	1	0,32	0,01					0
50	1	0,18	0,02					0
87	1	0,69	0,01					0
41	1	0,14	0,03					0

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Classé selon	Probabilité de sélection par niveau					Numéro des unités finales sélectionnées	
	Original				Final	Special	Si dans l'échantillon	Valeur espérée
Sno	ci	f0*	f1*	f2	=f	Pondérations (D)	bi	bi*D*f1
77	1	0,43	0,01					0
39	2	0,14	0,07					0
90	2	0,77	0,01					0
81	2	0,48	0,02					0
40	3	0,14	0,08					0
5	3	0,04	0,27					0
43	3	0,14	0,08					0
63	3	0,28	0,05					0
19	4	0,07	0,21					0
8	5	0,04	0,47	1,00	0,019	2,05	5	5
16	5	0,07	0,30	1,00	0,020	1,99	5	3
2	5	0,04	0,51	1,00	0,021	1,89	5	5
15	6	0,06	0,40	1,00	0,024	1,60	6	4
48	7	0,16	0,18	1,00	0,028	1,40	7	2
29	7	0,10	0,28	1,00	0,029	1,36	7	3
82	7	0,48	0,06	1,00	0,029	1,35	7	1
75	8	0,41	0,07	1,00	0,030	1,30	8	1
21	8	0,08	0,41	1,00	0,032	1,20	8	4
80	9	0,47	0,07	1,00	0,033	1,16	9	1
1	9	0,04	0,85	1,00	0,034	1,15	9	9
3	9	0,04	0,85	1,00	0,034	1,14	9	9
99	9	1,00	0,03	1,00	0,035	1,12	9	0
33	9	0,11	0,31	1,00	0,036	1,09	9	3
7	12	0,04	1,00	0,97	0,039	1,00	12	12
69	11	0,35	0,12	0,94	0,039	1,00	10	1
32	11	0,11	0,40	0,92	0,039	1,00	10	4
18	11	0,07	0,61	0,92	0,039	1,00	10	6
36	11	0,13	0,34	0,91	0,039	1,00	10	3
38	11	0,13	0,33	0,89	0,039	1,00	10	3
10	18	0,04	1,00	0,89	0,039	1,00	16	16
12	11	0,05	0,87	0,88	0,039	1,00	10	9
37	12	0,13	0,35	0,85	0,039	1,00	10	4
53	15	0,19	0,31	0,65	0,039	1,00	10	3
56	16	0,21	0,29	0,63	0,039	1,00	10	3
64	16	0,31	0,20	0,61	0,039	1,00	10	2
20	17	0,08	0,86	0,60	0,039	1,00	10	9
27	18	0,09	0,75	0,56	0,039	1,00	10	8
51	20	0,18	0,42	0,51	0,039	1,00	10	4
46	20	0,15	0,52	0,51	0,039	1,00	10	5
24	20	0,09	0,90	0,51	0,039	1,00	10	9
72	20	0,38	0,21	0,50	0,039	1,00	10	2

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Classé selon	Probabilité de sélection par niveau					Numéro des unités finales sélectionnées	
	Original				Final	Special	Si dans l'échantillon	Valeur espérée
Sno	ci	f0*	f1*	f2	=f	Pondérations (D)	bi	bi*D*f1
59	20	0,23	0,35	0,49	0,039	1,00	10	4
57	21	0,22	0,37	0,47	0,039	1,00	10	4
52	22	0,19	0,44	0,46	0,039	1,00	10	4
68	22	0,34	0,25	0,45	0,039	1,00	10	3
25	27	0,09	1,00	0,45	0,039	1,00	12	12
26	23	0,09	0,99	0,44	0,039	1,00	10	10
71	23	0,37	0,24	0,43	0,039	1,00	10	2
28	23	0,10	0,94	0,43	0,039	1,00	10	10
58	25	0,23	0,42	0,41	0,039	1,00	10	4
88	26	0,72	0,14	0,39	0,039	1,00	10	1
61	27	0,25	0,42	0,38	0,039	1,00	10	4
35	53	0,12	1,00	0,32	0,039	1,00	17	18
44	35	0,14	0,93	0,29	0,039	1,00	10	9
42	61	0,14	1,00	0,28	0,039	1,00	17	17
85	36	0,61	0,23	0,28	0,039	1,00	10	2
45	51	0,15	1,00	0,26	0,039	1,00	13	14
100	39	1,00	0,15	0,26	0,039	1,00	10	2
70	40	0,36	0,42	0,25	0,039	1,00	10	4
47	41	0,16	**1,00**	0,25	0,039	1,00	10	10
79	42	0,46	0,35	0,24	0,039	1,00	10	4
49	74	0,17	**1,00**	0,23	0,039	1,00	17	18
67	51	0,33	0,61	0,20	0,039	1,00	10	6
54	93	0,20	1,00	0,19	0,039	1,00	18	18
55	73	0,21	1,00	0,19	0,039	1,00	14	14
76	55	0,43	0,50	0,18	0,039	1,00	10	5
60	68	0,23	1,00	0,17	0,039	1,00	12	12
96	60	1,00	0,23	0,17	0,039	1,00	10	2
65	117	0,31	1,00	0,12	0,039	1,00	14	15
94	92	1,00	0,35	0,11	0,039	1,00	10	4
73	190	0,39	1,00	0,10	0,039	1,00	19	19
78	107	0,46	0,90	0,09	0,039	1,00	10	9
84	268	0,60	1,00	0,06	0,039	1,00	17	18
89	162	0,74	0,85	0,06	0,039	1,00	10	9
86	220	0,66	1,00	0,06	0,039	1,00	13	13
92	382	0,85	1,00	0,05	0,039	1,00	17	18
93	292	1,00	1,00	0,04	0,039	1,00	11	12
97	383	1,00	1,00	0,04	0,039	1,00	15	15
Total								500

Chapitre 6
Illustration du plan d'échantillonage et des procédures de sélection

6.1. Introduction

Le présent chapitre part d'une enquête réelle sur le travail des enfants dans un pays en développement pour illustrer en termes numériques détaillés les aspects du plan d'échantillonage et des procédures de sélection. Nous espérons qu'il permettra de clarifier et d'élargir la compréhension de ces procédures.

Comme celà est souvent le cas, l'enquête totale sur le travail des enfants implique deux composantes conceptuellement distinctes. Si l'on s'en tient à la terminologie précedemment présentée, il s'agit: (1) d'une *enquête sur le travail des enfants* (ETE) dont l'objectif premier est de mesurer l'ampleur du travail des enfants, c'est-à-dire la proportion des enfants travailleurs selon certaines définitions spécifiques des termes; et (2) d'une enquête sur les enfants travailleurs (EET) destinée surtout à la recherche des causes profondes, des caractéristiques, des circonstances et des conséquences du travail des enfants. Ces deux composantes sont liées entre elles et sont également souvent associées à deux autres opérations, à savoir: (3) L'Enquête sur la Main d'œuvre (EMO) destinée à évaluer l'activité économique de la population totale en âge de travailler; et (4) l'opération de listage des ménages dont l'objectif premier est de créer une base de sondage des ménages ou des unités semblables au sein des zones d'échantillon.

Ces composantes peuvent être liées de plusieurs façons et divers types de liens substanciels et opérationnels sont possibles. Certains peuvent avoir plus d'un objectif et/ou peuvent être associés à d'autres opérations, en fonction des circonstances et des exigences spécifiques du système d'enquête. Une disposition courante est la séquence

Listage—EMO—ETE—EET

où les composantes de la séquence impliquent progressivement des tailles d'échantillon plus petites et une collecte de données plus intensive et détaillée. Ce type de séquence d'enquêtes, y compris certaines de ses variations, souligne l'exposé des chapitres 4 et 5.

Dans ce chapitre, nous fournirons une illustration détaillée du plan de sondage et des procédures pour une disposition alternative de ces composantes, qui est possible et appropriée dans certaines circonstances. Cette disposition comporte deux phases: La phase 1, constituée d'une opération combinée Listage-ETE sur un échantillon relativement vaste; ensuite, la phase 2 combinant les opérations EET-EMO sur un échantillon substantiellement plus petit. Ce type de disposition se dessine d'elle-même dans les situations où ni les listes de ménages ni une enquête majeure comme l'EMO ne sont disponibles pour y fonder l'échantillon d'ETE; et par ailleurs, lorsque l'occasion de

conduire une enquête sur le travail des enfants peut être utilisée pour combler – ne serait-ce qu'en partie – le manque d'information sur la main d'œuvre générale.

Une différence majeure devrait être relevée entre le modèle EET abordé au chapitre 5 et le modèle de la phase 2 actuellement examiné dans ce chapitre. Dans le premier cas, l'objectif était d'obtenir un échantillon assez représentatif d'enfants travailleurs dans les ménages. Par conséquent, les ménages qui n'avaient pas d'enfant travailleur ne constituaient pas la principale cible. En outre, pour des raisons pratiques et économiques, on a envisagé d'exclure complètement de l'échantillon d'EET les zones n'ayant aucun ménage comptant des enfants travailleurs ou même celles n'ayant qu'un nombre réduit de tels ménages. En revanche, dans cette illustration, la phase 2 est requise pour produire des données non seulement sur les enfants travailleurs, mais aussi sur les caractéristiques de la main d'œuvre de la population générale. L'échantillon doit donc rester représentatif de la population générale. Des zones – même si, d'après la phase précédente, elles ne contiennent aucun enfant travailleur – doivent avoir une probabilité non nulle de faire partie de l'échantillon de la phase 2.

L'illustration spécifique des sections suivantes est fondée sur certaines propositions d'une enquête nationale sur le travail des enfants prévue au Yémen. Les données proviennent d'une base de sondage existante des zones de dénombrement du recensement (ZD)[32]; le contexte et les objectifs de l'enquête s'appuient sur la situation réelle. Mais il convient de relever qu'il ne s'agit pas d'un plan déjà mis en œuvre et que l'enquête, quand elle sera menée, peut utiliser un plan quelque peu différent.

Le plan d'échantillonnage doit bien sûr commencer par la définition du contexte et des objectifs de l'enquête. Celà est fourni dans la section 6.2 pour l'enquête illustrée. Cette section fournit également un résumé des principales caractéristiques d'un plan recommandé, lesquelles sont développées avec des détails techniques dans les sections suivantes.

Les sections 6.3 à 6.5 présentent respectivement les procédures d'échantillonnage pour:

- Le listage et l'ETE dans la phase 1
- L'EET dans la phase 2 et
- d'autres aspects de l'échantillonnage dans la phase 2, surtout pour le module EMO.

Dans la section 6.6, nous construisons une base de sondage hypothétique ayant des caractéristiques semblables à celles de la population étudiée et nous illustrons avec des chiffres l'application de diverses procédures d'échantillonnage développées dans les sections précédentes. La grande partie de cette illustration ne nécessite pas la sélection d'échantillons particuliers. Une sélection d'échantillon réelle est illustrée dans la section 6.7.

[32] Nous remercions le gouvernement du Yémen d'avoir mis ces informations à notre disposition.

6.2. Contexte et objectifs de l'enquête dans l'illustration

Comme souligné, cette illustration est basée sur certaines recommandations élaborées pour une enquête nationale sur le travail des enfants programmée. Pour la convenance de la présentation, nous supposerons que l'enquête a été programmée et qu'elle doit être réalisée comme elle est décrite.

L'enquête est censée être menée en deux phases:

- Phase 1, composante Listage-ETE. Listage des ménages et mesure de l'ampleur du travail des enfants dans les zones d'échantillon.
- Phase 2, composante EET-EMO. La conduite d'entretiens détaillés sur les enfants travailleurs et la main d'œuvre sur l'ensemble de la population en âge de travailler.

Pour la phase 1, les parties urbaines et rurales de chaque gouvernorat demandent les résultats. Le pays est constitué de 21 gouvernorats regroupés en 41 domaines de rapport.

6.2.1. Objectifs

1. Conduire une enquête à grande échelle pour l'estimation de l'ampleur du travail des enfants au niveau du gouvernorat séparément pour les zones urbaines et rurales.

2. Fournir des enquêtes détaillées sur les enfants travailleurs, dans les ménages identifiés comme ayant de tels enfants pendant la phase I.

3. Identifier les enfants travailleurs et enquêter en détail sur ce phénomène, dans d'autres ménages, même s'ils n'ont pas été identifiés comme ayant de tels enfants pendant la phase I. De tels enfants sont censés exister principalement (et en grand nombre) dans les ménages ayant des enfants d'âge scolaire mais non scolarisés. Certains – mais probablement en plus faible proportion – enfants travailleurs peuvent aussi exister dans d'autres ménages.

4. Obtenir l'information sur un «groupe de contrôle» d'enfants non travailleurs, nécessaire à une analyse plus approfondie des causes, des caractéristiques, des circonstances et des conséquences du travail des enfants.

5. Profiter de l'enquête pour collecter également les données sur les caractéristiques de la main d'œuvre de l'ensemble de la population en âge de travailler (le module EMO).

6.2.2. Principales caractéristiques du plan d'échantillonnage

Dès le départ, il est utile de résumer brièvement les principales caractéristiques du plan d'échantillonnage proposé.

1. Pour la phase 1, un grand échantillon est requis. Toutefois, la taille d'échantillon doit être limitée pour permettre le contrôle de la qualité des données collectées. Tous les ménages se trouvant dans les ZD seront énumérés et l'information glanée

sur la présence de travail des enfants. On recommande une taille de l'échantillon ne dépassant pas 800 ZD, soit un total de 100 000 ménages pour le listage et l'identification des enfants travailleurs.

2. L'échantillon sera réparti entre les 41 domaines de rapport, notamment la partie urbaine et rurale de chaque gouvernorat, de telle sorte que les plus grands échantillons aient de grands domaines, mais il n'est pas question de répartition en proportion de la taille du domaine. La répartition est un compromis entre les objectifs consistant à produire des estimations au niveau national ou au niveau du domaine. A quelques exceptions près, une certaine taille d'échantillon minimale est ventilée à chaque domaine, indépendamment de sa taille.

3. La base de sondage des ZD sera stratifiée suivant le type de lieu et la circonscription administrative ainsi que suivant la taille de la population de la zone et éventuellement d'autres critères pertinents dans la mesure du possible. Au sein de chaque strate, un échantillon à probabilité égale de ZD sera prélevé indépendamment en recourant à la procédure d'échantillonnage systématique.

4. La phase 2 sera réalisée sur un sous-échantillon des ZD de la phase 1. Dans les zones sélectionnées, les listes de ménages telles qu'elles sont préparées dans la phase 1 fourniront la base de sondage pour la sélection des ménages. La phase 2 utilise donc un plan à deux degrés.

5. Un sous-échantillonnage des ZD de la phase 1 à la phase 2 sera à probabilité proportionnelle aux mesures de la taille ainsi définies de manière à élever la probabilité de sélection des ZD là où le travail des enfants est, ou est susceptible d'être, élevé d'après les résultats de la phase 1. L'échantillon sera ainsi géographiquement concentré pour l'enquête sur les enfants travailleurs.

6. Un temps de décalage approprié entre les deux étapes est essentiel lors d'une telle procédure, pour que la base de sondage nécessaire puisse être compilée à partir de la phase 1 pour la sélection de l'échantillon de la phase 2.

7. La liste des ménages dans chaque ZD sélectionnée pour la phase 2 sera divisée en deux catégories: la catégorie C, ménages ayant un ou plusieurs enfants travailleurs ou des enfants non scolarisés; et la catégorie E, d'autres ménages. Les ménages de la catégorie (C) ont une forte probabilité de compter des enfants travailleurs, alors que ceux de la catégorie (E) ont une probabilité faible.

8. On recommande que la taille de l'échantillon n'excède pas 6 000 ménages de la catégorie C et 4 000 autres ménages de la catégorie E. Ils peuvent provenir de 350 à 400 ZD. Les résultats constituent le principal centre d'intérêt pour divers groupes de population au niveau national avec une répartition en fonction du critère urbain/rural et selon le sexe ou par principale division géographique du pays.

9. Tous les ménages de la catégorie C seront inclus dans l'échantillon jusqu'à un certain nombre maximal (25 à 35). Il s'agira du nombre sélectionné si le nombre disponible dans cette zone dépasse cette limite.

10. En ce qui concerne les ménages de la catégorie E de la même ZD, le taux d'échantillonnage variera inversement à la probabilité de sélection des ZD dans la

phase 1, donnant lieu aux limites inférieures et supérieures du nombre de tels ménages sélectionnés dans chaque zone. Cette procédure rendra uniformes, au sein des domaines, les probabilités de sélection des ménages de la catégorie E et moins variable le nombre de ménages sélectionnés.

Les trois sections suivantes 6.3 à 6.5 décrivent respectivement:

- L'échantillonnage pour le listage et l'ETE dans la phase 1.
- L'échantillonnage pour l'EET dans la phase 2.
- D'autres aspects de l'échantillonnage, surtout du point de vue des exigences des composantes EMO dans la phase 2.

Avant d'examiner les questions liées au plan d'échantillonnage, il est nécessaire de souligner certains aspects de la situation du travail des enfants au Yémen relatifs à la sélection du plan.

6.2.3. Certains aspects pertinents de la situation du travail des enfants au Yémen

Un rapport d'évaluation rapide[33] sur le travail des enfants fournit la description suivante[34]:

- Le travail des enfants est un problème grave au Yémen, et bien plus encore, il semble que ce mal prend de l'ampleur assez rapidement.
- Parmi les enfants de 6 à 14 ans, plus d'un tiers (37 pour cent) ne vont pas à l'école. Cela concerne près d'un garçon sur cinq (18 pour cent) et plus de la moitié (55 pour cent) des filles âgées de 6 à 14 ans.

En termes réels, les chiffres ci-dessus s'élèvent à 2 millions d'enfants non scolarisés sur un total de 5,4 millions parmi les 6 à 14 ans. Lors d'une enquête sur la main d'œuvre qui couvrait aussi le travail des enfants, sensiblement 0,9 millions de ces enfants ont été signalés comme travailleurs et les 1,1 millions restants ont été signalés comme non scolarisés. De plus, ce dernier groupe semble être constitué entièrement de filles; la plupart des garçons non scolarisés semblent être signalés comme travailleurs.

Cette description semble se présenter plus ou moins de la manière suivante:

Nombre d'enfants âgés de 6 à 14 ans (en millions)					
	Total	Scolarisés	Non scolarisés	Parmi les non scolarisés	
				travailleurs	non travailleurs
Garçons	2,7	2,2	0,5	0,4	0,1
Filles	2,7	1,2	1,5	0,5	1,0
Tous	5,4	3,4	2,0	0,9	1,1

Il est très probable qu'une bonne partie de l'activité économique des filles soit perçue à tort comme «tâches domestiques» et ne soit pas déclarée comme «travail» dans les

[33] Shaik, Khaled Rajeh. 2000-2001. *Child Labour in Yemen* (Sana'a, Ministry of Labour of Yemen).

[34] Tous les chiffres avancés dans ce rapport sont approximatifs et certaines statistiques signalées sont incohérentes. Ils fournissent néanmoins une vue d'ensemble assez claire.

enquêtes. Un objectif important de l'enquête sur le travail des enfants devrait consister à identifier et à quantifier ce phénomène.

On peut relever que, de cette façon, 90 pour cent des enfants travailleurs (c'est-à-dire les enfants déclarés comme «travailleurs») travaillent pour leurs familles sans rémunérations – cela s'applique aussi bien aux garçons qu'aux filles. La plupart (plus de 90 pour cent) de ces activités sont concentrées dans l'agriculture et les secteurs voisins, et par conséquent dans les zones rurales. (L'ampleur du travail des enfants peut être 5 à 7 fois plus élevée en zones rurales qu'en ville). Si les garçons et les filles forment des proportions identiques (50 pour cent chacun) d'enfants déclarés comme travailleurs, les filles représentent 55 pour cent du travail dans l'agriculture et les secteurs voisins, tandis que les garçons représentent 70 à 95 pour cent dans les principaux secteurs non agricoles comme le commerce, les services, l'artisanat, le travail non qualifié et surtout le travail dans la rue.

Le travail des enfants semble représenter un sixième environ de la main-d'œuvre totale!

On peut aussi souligner quelques faits démographiques importants pour le plan d'échantillonnage. Avec une population totale d'environ 20 millions, il y a 2,7 millions de ménages dans le pays, ce qui donne une taille moyenne des ménages de 7,4 personnes. La main d'œuvre totale est de l'ordre de 5,5 millions, ce qui donne une moyenne de 2 personnes économiquement actives par ménage. En moyenne, il y a 0,35 enfants travailleurs (âgés de 5 à 14 ans) par ménage et 0,40 autres enfants non scolarisés. Cela veut dire par exemple que dans un listage d'échantillon de 75 000 ménages, on peut compter 15 000 à 20 000 ménages ayant un enfant qui travaille, 20 000 à 25 000 ayant un enfant non scolarisé (mais non déclaré travailleur), et 30 000 à 40 000 dont tous les enfants sont scolarisés ou n'ayant pas d'enfants dans le groupe d'âge concerné.

6.3. Echantillonnage pour listage et ETE

6.3.1. Introduction

Les opérations de listage et d'ETE de la phase 1, sont concernés par le premier objectif majeur de l'enquête planifiée nommée ci-dessus, à savoir, conduire une enquête à grande échelle pour l'estimation de l'ampleur du travail des enfants au niveau du gouvernorat, séparément pour les zones urbaines et rurales.

Cela nécessite une enquête sur une grande échelle, en raison du grand nombre (41) de domaines pour lesquels il faut des estimations distinctes, suffisament fiables. Cela necessite d'utiliser des questions brèves et très simples dans l'enquête. Cependant, le problème majeur d'une telle opération est la sous-déclaration du travail des enfants, qui peut être réduite en recourant à des questions bien réfléchies et ciblées.

Objectifs de l'enquête et taille d'échantillon

Un compromis est requis entre une taille d'échantillon grande d'une part, et une mesure plus exacte des cas inclus, d'autre part.

Avant tout, le compromis implique une nouvelle appréciation des objectifs formulés pour produire des estimations pour environ 41 domaines, la plupart étant assez petits en termes de taille de la population. Pour ce qui est de l'enquête de listage à utiliser pour mesurer le niveau du travail des enfants, il est probable qu'elle soit subordonnée à un biais considérable de sous-déclaration. Il n'est pas indiqué de choisir une taille d'échantillon sur la seule base des exigences de *précision de l'échantillonnage*. Par contre, la principale préoccupation devrait être d'obtenir des estimations fiables premièrement au niveau national, puis par critère urbain/rural et selon le sexe, éventuellement ensuite suivant les quatre ou cinq principales régions géographiques du pays. Les estimations des plus grands gouvernorats constitueront l'étape suivante. A ce stade, une décomposition accrue ne peut être qu'à titre indicatif, pour trouver de très grandes différences. Il est aussi possible d'utiliser certaines méthodes «synthétiques» simples pour une décomposition des résultats avec plus de fiabilité que sur la seule base des estimations purement «directes» tirées des échantillons de taille relativement petite. (Voir Section 7.5 pour un exposé préliminaire de telles procédures d'estimation).

Certains de ces points seront détaillés davantage dans les sections suivantes.

Evaluation de l'ampleur du travail des enfants pendant le listage

Au Yémen, de nombreux enfants ne vont pas à l'école et cependant ils ne sont pas déclarés dans les enquêtes comme étant engagés dans une forme d'activité économique ou de travail des enfants. Cela s'applique surtout dans le cas des filles. Cette situation est certainement causée par la sous-déclaration générale du travail des enfants.

Il ne suffit pas de poser une simple question comme par exemple «Y-a-t-il un enfant qui travaille dans le ménage?». Il faut au moins inclure une petite série de questions dans le questionnaire pour classer les ménages en fonction des types suivants:

1. Ménages déclarant un ou plusieurs enfants travailleurs.
2. Aucun enfant travailleur n'est déclaré, mais le ménage déclare un ou plusieurs enfants en âge scolaire qui ne vont pas à l'école, c'est-à-dire ne faisant pas d'études à temps complet.
3. Tous les enfants sont déclarés scolarisés (étude à temps complet) et aucun n'est déclaré au travail?
4. Le ménage n'a aucun enfant dans le groupe d'âge concerné.

Comme décrit plus loin, pour l'enquête sur les enfants travailleurs de la phase 2 (en plus du module EMO), les groupes (1) et (2) – appelés ménages de la «catégorie C» – devront être échantillonnés différemment des groupes (3) et (4) appelés ici ménages de la «catégorie E».

L'ensemble minimum de questions permettant d'identifier la catégorie ci-dessus pourrait ressembler à celui-ci:

- Combien d'enfants (du groupe d'âge concerné) y a-t-il dans le ménage?
- Combien d'entre eux vont à l'école à temps complet et combien ne vont pas à l'école ou ne font pas des études à temps complet?

- Combien de ceux du second groupe (non scolarisés) sont engagés dans le travail, y compris dans le travail non payé de l'activité économique du ménage?
- Certains enfants du premier groupe (scolarisés) sont-ils engagés dans un tel travail?

Cela implique que l'opération de listage ne peux (et ne doit) pas être superficielle; elle doit produire des données fiables sur les points ci-dessus. Par conséquent, il est très important de ne pas trop étendre l'opération en termes de nombre de ménages devant être listés. L'opération devrait ressembler davantage à un «mini entretien».

6.3.2. Base de sondage

Exigences de base

Les zones de dénombrement (ZD) du recensement formeront les unités d'échantillonnage primaires (UEP). La base de sondage devrait contenir au moins les informations suivantes:

- Liste des ZD couvrant tout le pays, assortie de cartes et d'une description.
- Les codes identifiant les unités géographiques/administratives (gouvernorat, sous-gouvernorat, etc.) auxquelles la ZD appartient, selon qu'elle est urbaine, rurale ou nomade.
- La taille de la population et, à souhait, le nombre de ménages dans la ZD aussi.

Si elles sont disponibles, d'autres caractéristiques potentiellement utiles pour une stratification plus efficace peuvent aussi être intégrées. Une variable particulière – la classification des ZD en groupes étendus en termes de taux moyen d'analphabétisme – devrait être mentionnée comme une possibilité.

Chiffres de bases

Dans la sélection de la base de sondage, il est utile de garder en vue le nombre d'unités (ZD, ménages, personnes) impliquées et leurs caractéristiques de base, telles que leur taille moyenne. Ces chiffres devraient être répartis par circonscriptions administratives et selon le critère urbain/rural, etc.

A partir de la base de sondage disponible, le tableau 6.2 présente des chiffres de base par gouvernorat et par type de zone (urbaine, rurale, nomade). Des tableaux plus détaillés sont disponibles pour les circonscriptions administratives plus petites et bien sûr pour chaque ZD de la base de sondage complète.

Tableau 6.1: Taille des principaux domaines

	Urbain	Rural	Nomade	Total
% du total dans le pays				
Personnes	28,2	71,3	0,5	100,0
M	29,1	70,4	0,5	100,0
ZD	25,6	72,4	2,0	100,0
Taille relative des ZD (=1,0 pour tout le pays)				
M/ZD	1,14	0,97	0,24	1,00

Premièrement, le tableau 6.1 résume la distribution de la population par type de zone. Les zones urbaines représentent 28 pour cent de la population, mais les zones nomades 0,5 pour cent seulement. Les ZD urbaines sont quelque peu plus grandes que les ZD rurales en termes de population (sensiblement près de 17 pour cent), mais les ZD nomades sont nettement plus petites – soit en moyenne un quart de la taille des autres ZD.

Tableau 6.2: Nombre de personnes, de ménages et de ZD en fonction du critère urbain, rural et du gouvernorat (à partir de la base de sondage du recensement)

	Gouver.	Personnes	M	ZD	Personnes/M	M/ZD
URBAIN	11	368 726	51 339	371	7,2	138
	12	109 514	15 566	106	7,0	147
	13	1 654 933	244 693	1637	6,8	149
	14	105 551	13 459	91	7,8	148
	15	528 601	80 349	621	6,6	129
	16	55 232	7 687	44	7,2	175
	17	132 934	17 493	135	7,6	130
	18	748 697	111 903	803	6,7	139
	19	468 542	59 049	428	7,9	138
	20	182 272	24 766	183	7,4	135
	21	73 729	8 670	74	8,5	117
	22	101 605	13 666	91	7,4	150
	23	25 103	3 483	29	7,2	120
	24	567 133	89 695	633	6,3	142
	25	60 150	9 057	65	6,6	139
	26	29 700	3 760	29	7,9	130
	27	35 119	4 597	35	7,6	131
	28	35 241	5 468	40	6,4	137
	29	146 308	19 117	152	7,7	126
	30	60 081	8 196	62	7,3	132
	31	3 652	556	5	6,6	111

	Gouver.	Personnes	M	ZD	Personnes/M	M/ZD
RURAL	11	1 752 020	250 331	2 057	7,0	122
	12	311 994	40 979	356	7,6	115
	13	39 994	4 647	41	8,6	113
	14	472 696	53 205	441	8,9	121
	15	1 860 061	284 078	2 286	6,5	124
	16	365 511	48 300	290	7,6	167
	17	1 327 485	175 416	1 503	7,6	117
	18	1 397 351	235 482	1 878	5,9	125
	19	533 210	60 892	537	8,8	113
	20	1 136 526	161 502	1 416	7,0	114
	21	383 309	41 216	397	9,3	104
	22	580 991	70 650	608	8,2	116
	23	885 392	112 865	913	7,8	124
	24					
	25	654 953	93 758	756	7,0	124
	26	202 490	23 744	208	8,5	114
	27	457 181	63 951	531	7,1	120
	28	36 835	5 732	67	6,4	86
	29	720 503	86 269	753	8,4	115
	30	411 296	51 120	409	8,0	125
	31	385 959	54 780	480	7,0	114
TOTAL URBAIN		5 492 823	792 569	5 634	6,9	141
TOTAL RURAL		13 915 757	1 918 917	15 927	7,3	120
TOTAL NOMADE		96 492	13 397	443	7,2	30
TOTAL DU PAYS		19 505 072	2 724 883	22 004	7,2	124

La variation de la taille des ZD est un facteur déterminant indiquant la procédure appropriée pour choisir un échantillon de ces unités. Le tableau 6.3 fournit les indicateurs sur la variabilité de la taille des ZD, tout d'abord parmi les moyennes des gouvernorats et ensuite parmi les moyennes des entités administratives plus petites. Dans ce dernier cas, une bonne partie de la variabilité est représentée par environ 20 sous-gouvernorats à chaque fin de distribution.

La variabilité de la taille de la population de chaque ZD est bien sûr susceptible d'être plus considérable que celle indiquée par les chiffres susmentionnés dont on a fait les moyennes sur les circonscriptions administratives. On peut examiner cela à partir de la base de sondage complète fournissant les listages au niveau des ZD.

Tableau 6.3: Nombre moyen de ménages par ZD: variation par gouvernorat			
	Urbain	Rural	Nomade
Nombre de gouvernorats	21	20	15
Nombre moyen de M par ZD			
Valeur minimale	111	86	1
Valeur maximale	175	167	69
Moyenne sur les gouvernorats	136	119	25
Valeur médiane	137	116	22
Coefficient de variation (cv)	0,10	0,12	0,74
Nombre moyen de ménages par ZD: variation par sous-gouvernorat			
	Urbain	Rural	Nomade
Nombre de sous-gouvernorats			
Nombre moyen de M par ZD	239	309	88
Valeur minimale	51	12	1
Valeur maximale	298	311	151
Moyenne sur les sous-gouvernorats	134	119	30
Valeur médiane	134	120	21
Coefficient de variation (cv)	0,22	0,22	0,92
Indicateurs ci dessus, en ommettant les 20 premières valeurs et les 20 dernières de la distribution			
Nombre moyen de M par ZD			
Valeur minimale	100	88	
Valeur maximale	164	137	
Moyenne sur les sous-gouvernorats	133	119	
Valeur médiane	134	120	
Coefficient de variation (cv)	0,12	0,08	

6.3.3. Taille de l'échantillon et répartition

Dans une enquête nationale, la plus importante considération pratique est la taille totale de l'échantillon au niveau du pays – indépendamment des exigences de déclaration des domaines sous-nationaux.

Les listages, et par ricochet les questions de base de l'ETE, doivent couvrir tous les ménages dans chaque ZD sélectionnée – le sous-échantillonnage des ménages au sein d'une ZD n'est pas une option dans une opération[35] de listage. De très grandes ZD peuvent d'abord être segmentées avant le listage et l'opération de listage peut se limiter à un échantillon d'un ou plusieurs segments de la ZD. Toutefois, cette opération peut être coûteuse et difficile, bien qu'elle soit néanmoins parfois nécessaire.

En raison du prélèvement de tous les ménages de chaque zone sélectionnée dans l'échantillon, le nombre de ménages de l'échantillon est directement déterminé par le nombre de ZD sélectionnées et par la taille moyenne de la ZD.

Le tableau 6.4 indique que la taille maximale de l'échantillon recommandée est 100000 ménages à sonder sur le terrain dans la phase 1. Pour la répartition particulière présumée ici, le nombre correspondant de ZD dans l'échantillon sera d'environ 870. Il

[35] Le sous-échantillonnage est bien sûr possible pour l'ETE si l'ETE est opérationnellement séparée du listage des ménages. Cela peut être significatif pour une enquête plus approfondie que l'ETE présente impliquant seulement une poignée de questions pour le triage.

peut être souhaitable de rendre l'échantillon quelque peu plus petit encore – de 20 pour cent par exemple- étant donné le besoin d'insister sur la qualité des données comme précédemment relevé.

Domaines de rapport

Un total de 41 domaines de rapport a été indiqué: les zones urbaines de chacun des 21 gouvernorats et les zones rurales de 20 gouvernorats qui ont de telles zones.

Il est nécessaire d'affiner un peu cette spécification.

Tout d'abord, toutes les zones nomades devraient être perçues comme formant un domaine de rapport important à leur propre compte. (La répartition des zones nomades par gouvernorat n'est pas possible en raison de leur petite taille).

Deuxièmement, certains domaines peuvent être si petites qu'une taille d'échantillon suffisante ne peut leur être raisonnablement affectée pour une estimation séparée. Cela est particulièrement vrai pour le secteur urbain du gouvernorat 31 (voir tableau 6.2 ci-dessus). Si de très petits groupes de ce genre ne peuvent pas de façon réaliste constituer des domaines pour une déclaration séparée, il est généralement convenable et approprié de les conserver comme des strates distinctes aux fins de la sélection d'échantillon.

Répartition de l'échantillon

Pour les niveaux semblables de précision, différents domaines – indépendamment de leurs tailles, nécessitent des tailles d'échantillon égales. Par contre, pour la précision minimale des estimations au niveau national, la répartition requise est presque proportionnée. Aucun de ces extrêmes n'est indiquée lorsque l'objectif consiste à produire à la fois des estimations au niveau du domaine et au niveau national.

Un compromis courant et approprié consiste à répartir l'échantillon proportionnellement à la racine carrée de la taille de la population, c'est-à-dire comme

$$n_i = n_0 \cdot M_i^{1/2}$$

où $M_i = (P_i/)$ est la taille relative de la population du domaine i, c'est-à-dire sa population divisée par la population moyenne par domaine; n_0 est une constante déterminée pour obtenir la taille totale de l'échantillon cible n, autrement dit

$$n_0 = n/\Sigma M_i^{1/2}$$

Pour assurer une taille d'échantillon minimale pour chaque domaine, y compris les plus petits, l'expression ci-dessus peut être généralisée comme suit.

$$n_i = n_0[k^2 + (1-k^2)M_i^{\alpha}]^{1/2}$$

K est ici un facteur déterminé par l'importance relative accordée à l'estimation au niveau du domaine par opposition au niveau national. Elle détermine la taille d'échantillon minimale de tout domaine, à savoir $n_{min} = n_0 \cdot k$ pour un domaine avec M_i proche de zéro.

Le paramètre α détermine jusqu'à quel point il est permis d'agrandir les vastes domaines. Une petite valeur de α fixe une limite plus précise sur l'échantillon maximal ventilé; cette limite augmente lorsque α s'accroit.

Le tableau 6.4 présente la répartition par domaine en prenant pour paramètres les valeurs suivantes.

k=0,5 ce qui équivaut à donner la même importance à ces deux objectifs de l'estimation au niveau du domaine et au niveau national.

α=1,0 qui équivaut à la répartition de la racine carrée la plus largement utilisée, à l'exception de la répartition minimale assurée par k

n=100.000 ménages, étant la taille d'échantillon maximale recommandée dans ce cas. Cela détermine la constante n_0 requise dans la formule[36] de répartition

Les divers domaines diffèrent dans une certaine mesure en termes de taille moyenne des ménages et de nombre de ménages ou de personnes par ZD. Pour enlever l'effet de cette variation, dans le tableau 6.4, la répartition a été réalisée de la même manière en termes de nombre de personnes. Les étapes ci-dessous clarifient la procédure appliquée.

1. Pour chaque domaine i, la mesure relative de la taille est définie en termes de nombre de *personnes* en son sein.

   $M_i = P_i/\overline{P}$

2. En termes de nombre de personnes, la répartition de la taille d'échantillon p_i est déterminée, en prenant une valeur relative (comme 1000) pour la mise à l'échelle du facteur n_0, et k=0,5, α=1.

3. Le taux d'échantillonnage $f_i = p_i/P_i$ ainsi obtenu est alors appliqué au nombre de ménages, H_i, du domaine pour obtenir le nombre escompté de ménages, h_i, de l'échantillon (correspondant à la valeur arbitraire du facteur d'ajustement n_0), $h_i = f_i \cdot H_i$.

4. Les taux d'échantillonnage f_i sont ajustés uniformément pour obtenir la taille d'échantillon cible, h_{cible}=100000 ménages.

   $f_i' = f_i \cdot h_{cible}/\Sigma h_i$

5. Enfin, le nombre a_i à sélectionner dans les ZD A_i du domaine est

   $a_i = f_i' \cdot A_i$

[36] Certains ajustements ont été faits dans le tableau 6.4 pour quelques domaines. Le domaine 31-urbain est si petit que l'allocation ci dessus signifierait un recencement complet. L'allocation pour le domaine nomade a été augmenté de 50% pour des raisons de pertinence.

6.3.4. Stratification et sélection des zones

Pour la sélection des zones, la stratification sera appliquée à divers niveaux.

1. Stratification par type de zone (urbaine, rurale, nomade).
2. Au sein de celles-ci, par gouvernorat.
3. Au sein de chaque gouvernerat, par niveau administratif immédiat.
4. Par groupe de taille de la population de la ZD.
5. Eventuellement suivies par la disposition des ZD par strates définies ci-dessus, par taille de la population, puis en appliquant l'échantillonnage systématique.

Est-il possible et souhaitable d'introduire une stratification supplémentaire?

La stratification des zones en fonction du niveau d'analphabétisme est une possibilité qui a été envisagée dans la zone. Utilisée, elle peut prendre la place de l'étape (4) ci-dessus, suivie éventuellement par (5), ordonnée par taille de la population de la ZD et par la sélection systématique.

Toutefois, cette option de stratification par niveau d'alphabétisme peut ne pas être possible en raison du manque de données pour chaque ZD. En outre, il n'est pas certain que ce sera une variable de stratification efficace. L'alphabétisme peut être un facteur qui fonctionne essentiellement au niveau individuel plutôt qu'au niveau de la zone. En tout cas, il est important de maintenir un bon contrôle sur la variation de la taille de la population de la ZD dans l'échantillonnage.

A l'intérieur de chaque strate, les ZD peuvent être sélectionnées avec des probabilités uniformes requises en recourant à la procédure systématique circulaire standard. (Voir la Section 6.7 pour une illustration numérique de cette procédure.)

Tableau 6.4: Illustration de la répartition de l'échantillon et des taux d'échantillonnage pour la sélection des ZD en fonction du critère urbain/rural et du gouvernorat (Taille d'échantillon cible: 100000 ménages)

		Nombre dans la population			Taux d'échantil-lonnage	Attendu	ZD
	Gouver.	Personnes	M	ZD	(f %)	n (M)	sélectionée
URBAIN	11	368 726	51 339	371	4,65	2 388	17
	12	109 514	15 566	106	11,13	1 732	12
	13	1 654 933	244 693	1637	1,93	4 715	32
	14	105 551	13 459	91	11,46	1 542	10
	15	528 601	80 349	621	3,71	2 979	23
	16	55 232	7 687	44	19,67	1 512	9
	17	132 934	17 493	135	9,57	1 673	13
	18	748 697	111 903	803	3,01	3 368	24
	19	468 542	59 049	428	3,99	2 359	17
	20	182 272	24 766	183	7,55	1 870	14
	21	73 729	8 670	74	15,37	1 333	11
	22	101 605	13 666	91	11,81	1 615	11
	23	25 103	3 483	29	37,42	1 303	11
	24	567 133	89 695	633	3,55	3 186	22
	25	60 150	9 057	65	18,27	1 655	12
	26	29 700	3 760	29	34,28	1 289	10
	27	35 119	4 597	35	29,42	1 352	10
	28	35 241	5 468	40	29,32	1 603	12
	29	146 308	19 117	152	8,89	1 700	14
	30	60 081	8 196	62	18,29	1 499	11
	31	3 652	556	5	37,42	208	2

		Nombre dans la population			Taux d'échantil-lonnage	Attendu	ZD
	Gouver.	Personnes	M	ZD	(f %)	n (M)	sélectionée
RURAL	11	1 752 020	250 331	2057	1,87	4 677	38
	12	311 994	40 979	356	5,19	2 127	18
	13	39 994	4 647	41	26,16	1 216	11
	14	472 696	53 205	441	3,97	2 114	18
	15	1 860 061	284 078	2286	1,81	5 139	41
	16	365 511	48 300	290	4,68	2 260	14
	17	1 327 485	175 416	1503	2,17	3 814	33
	18	1 397 351	235 482	1878	2,11	4 977	40
	19	533 210	60 892	537	3,69	2 246	20
	20	1 136 526	161 502	1416	2,37	3 828	34
	21	383 309	41 216	397	4,54	1 870	18
	22	580 991	70 650	608	3,50	2 473	21
	23	885 392	112 865	913	2,73	3 082	25
	24						
	25	654 953	93 758	756	3,26	3 054	25
	26	202 490	23 744	208	7,00	1 662	15
	27	457 181	63 951	531	4,06	2 594	22
	28	36 835	5 732	67	28,17	1 615	19
	29	720 503	86 269	753	3,08	2 656	23
	30	411 296	51 120	409	4,34	2 217	18
	31	385 959	54 780	480	4,52	2 474	22
NOMADE	total	96 492	13 397	443	22,59	3 026	**100**
TOTAL		19 505 072	2 724 883	22 004	3,67	100 000	869

Répartition avec des paramètres (voir texte) k=0,5, alpha=1.

Tailles et taux d'échantillonnage des principaux domaines

	Nombre dans la population			Attendu	ZD	moyenne
	Personnes	M	ZD	n (M)	Taux d'échantillonnage (%)	
URBAIN	5 492 823	792 569	5 634	40 882	297	5,158
RURAL	13 915 757	1 918 917	15 927	56 092	472	2,923
NOMADE	96 492	13 397	443	3 026	100	22,587
TOTAL	19 505 072	2 724 883	22 004	100 000	869	3,670

Taille relative et taux d'échantillonnage

	Nombre dans la population			Attendu	ZD	moyenne
	Personbes	M	ZD	n (M)	Taux d'échantillonnage	
URBAIN	28,2	29,1	25,6	40,9	34,1	1,41
RURAL	71,3	70,4	72,4	56,1	54,3	0,80
NOMADE	0,5	0,5	2,0	3,0	11,5	6,15
TOTAL	100,0	100,0	100,0	100,0	100,0	1,00

Les pondérations d'échantillon applicables au stade de l'estimation seront inversement proportionnelles aux taux d'échantillonnage f_i' définis dans le tableau 6.4 ci-dessus. On peut néanmoins relever une petite correction. Il en est ainsi parce que le nombre de ZD à sélectionner dans chaque strate est prévu pour être un nombre entier. Arrondir à l'entier le plus proche implique un léger ajustement du taux d'échantillonnage.

Soit f_i, le taux d'échantillonnage déterminé ci-dessus pour le domaine i. Soit J, indiquant une strate à l'intérieur du domaine i. D'autres niveaux de stratification tels que (3) et (4) ci-dessus signifient qu'un domaine peut renfermer plusieurs strates séparées.

Dans les ZD A_{ij} de la strate j du domaine i, le nombre à sélectionner est

$$a_{ij}=f_i \cdot A_{ij}$$

Soit a'_{ij}, la quantité arrondie à l'entier le plus proche qui est le nombre réel à sélectionner. Le taux d'échantillonnage effectif dans la strate est donc

$$f_{ij}=a'_{ij}/A_{ij}=f_i\ (a'_{ij}/a_{ij})$$

Les «pondérations du plan» sont l'inverse de ce f_{ij} qui est uniforme pour tous les ZD et les ménages sélectionnés dans la strate.

Dans l'exposé sur la phase 2 de la section suivante, nous nous référerons à ces taux d'échantillonnage comme les probabilités d'échantillonnage du premier stade ou de la phase 1 (l'identification explicite de la strate et du domaine n'est pas nécessaire pour cet exposé).

6.4. Echantillonnage pour l'enquête sur les enfants travailleurs (EET)

La détermination du plan d'échantillonnage implique les étapes suivantes:

1. Précision ou clarification des objectifs et des principales caractéristiques du plan. Elle est liée aux objectifs formels de l'enquête.
2. Préparation de la base de sondage, y compris la stratification, en affectant les mesures de la taille pour la sélection avec PPT si nécessaire et en estimant les paramètres dans la base de sondage, au besoin, pour le plan d'échantillonnage et la sélection. Cela concerne essentiellement le(s) stade(s) d'échantillonnage du domaine.
3. La sélection de la taille d'échantillon en termes d'unités d'échantillonnage finales, comme les ménages.
4. La précision des stades d'échantillonnage et le nombre de grappes (unités de zone finales) à sélectionner. Etant donné (3), elle détermine aussi le nombre moyen d'unités finales (ménages) par grappe.
5. La description de la méthode de sélection des unités de zones, en particulier les probabilités d'échantillonnage.
6. La description de la méthode de sélection des unités finales (ménages) au sein des grappes d'échantillon, en particulier les taux d'échantillonnage et les probabilités d'échantillonnage finales.

7. La précision de la pondération et de la procédure d'estimation, y compris l'estimation de la variance.

Elles sont exposées ci-dessous à leur tour.

6.4.1. Objectifs et principales caractéristiques du plan

Objectifs

On a précédemment relevé cinq objectifs majeurs de l'enquête planifiée. L'objectif (1), à savoir l'estimation de l'ampleur du travail des enfants par domaine urbain/rural de chaque gouvernorat est le centre d'intérêt de la phase 1 (listage et ETE) décrite ci-dessus. L'objectif de la phase 2 (EET), avec le «module» EMO a été relevé comme suit:

1. Les enquêtes détaillées sur les enfants travailleurs, dans les ménages identifiés comme ayant de tels enfants pendant la phase 1.

2. Les enquêtes détaillées sur les enfants travailleurs et leur identification, dans d'autres ménages, même s'ils n'ont pas été identifiés comme ayant de tels enfants pendant la phase 1. De tels enfants sont censés exister principalement (et en grand nombre) dans les ménages ayant des enfants d'âge scolaire mais non scolarisés. Certains – mais probablement en plus faible proportion – d'enfants travailleurs peuvent aussi exister dans d'autres ménages.

3. Obtenir l'information sur un «groupe de contrôle» des enfants qui ne travaillent pas, nécessaire à une analyse plus appronfondie des causes, des caractéristiques, des circonstances et des conséquences du travail des enfants.

4. Profiter de l'enquête pour collecter également les données sur les caractéristiques de la main d'œuvre de l'ensemble de la population en âge de travailler.

Etant donné le contexte de l'enquête entière, les objectifs (2) et (3) sont fondamentaux, tandis que les objectifs (4) et (5) – même s'ils sont importants – devraient être perçus comme secondaires. Ces priorités doivent également se refléter dans le plan d'échantillonnage. En d'autres termes, le plan doit être principalement défini pour atteindre les objectifs (2) et (3), mais s'accommoder dans la mesure du possible à (4) et (5). Parmi ces derniers, c'est l'objectif (5) qui définit principalement les exigences d'échantillonnage.

Principales caractéristiques

Par conséquent, d'un point de vue formel, la phase 2 peut être considérée comme ayant deux composantes: le point EET relatif à l'objectif (2) et (3); et la partie EMO concernant principalement l'objectif (5).

Bien évidemment, cette distinction conceptuelle n'implique pas une séparation opérationnelle. Les deux composantes prendront la forme d'un entretien unifié dans les mêmes zones d'échantillon et ménages, en appliquant les sections pertinentes du ou des questionnaire(s) dans chaque ménage. Toutefois, cette distinction est fondamentale du point de vue du plan de l'échantillon. Le déterminant primaire du plan

d'échantillonnage est la partie EET qui sera abordée en premier dans les sections suivantes.

L'entretien avec les enfants travailleurs (et/ou leurs tuteurs) est une opération intense et chronophage. Bien plus, un questionnement plus détaillé sera requis pour découvrir les enfants travailleurs qui ne sont pas faciles à identifier ou étaient manquants lors du listage et/ou de l'ETE pendant la phase 1. L'accent doit être mis sur la qualité des données et le contrôle des erreurs de non échantillonnage plutôt que sur une précision élevée de l'échantillonnage et une grande taille d'échantillon. Par ailleurs, le plan d'échantillonnage doit être aussi efficace que possible pour minimiser les tailles d'échantillon nécessaires pour un degré de précision donné.

Deuxièmement, pour les mêmes raisons évoquées ci-dessus, il est nécessaire à ce stade de se concentrer au niveau des estimations nationales. Les conditions des enfants travailleurs au Yemen sont peu connues. Certains des plus importants indicateurs peuvent être déclarés pour 3 à 5 principaux domaines au plus, par exemple en fonction du critère urbain/rural et genre ou par principales divisions géographiques du pays. Une répartition plus détaillée devrait attendre jusqu'à ce que les chiffres de base soient disponibles et suffisament fiables.

Troisièmement, l'échantillon devrait être concentré sur les zones (ZD) ayant une ampleur élevée de travail des enfants comme indiqué par les résultats de la phase 1. Sinon, l'échantillon sera dispersé sur les zones ayant très peu ou pas d'enfants travailleurs. Le nombre de zones d'échantillon dans la phase 2 devrait être plus petit que celui qui est dans l'échantillon de la phase 1. Cela est nécessaire pour veiller au contrôle de la qualité.

6.4.2. Base de sondage

Les ZD de l'échantillon et les listages de la phase 1 constituent la base de sondage de la phase 2. Dans les pages qui suivent, nous appellerons les premières la «population» et les dernières «échantillon» pour des raisons de commodité.

Il faudrait un décalage entre les deux phases pour que la base de sondage soit compilée pour la sélection de l'échantillon en phase 2. La base de sondage devrait contenir au moins les informations suivantes:

1. Liste des ZD assortie de cartes et d'une description pour leur localisation et leur identification.
2. Probabilités de sélection de la phase 1 (premier niveau d'échantillonnage) $f1_i$ de la ZD i. Elles varient par domaine et strate comme déjà relevé.
3. Liste de ménages dans les ZD, avec les informations relatives aux catégories suivantes, dans lesquelles chaque ménage appartient (pendant les listages de la phase 1):

La catégorie C: Le ménage déclare un ou plusieurs enfants travailleurs[37]; ou bien aucun enfant travailleur n'est déclaré, mais le ménage déclare un ou plusieurs enfants en âge scolaire qui ne vont pas à l'école (études à temps complet).

La catégorie E: Le ménage déclare que tous les enfants vont à l'école (études à temps complet) et qu'aucun ne travaille; ou bien qu'il n'y a aucun enfant dans la tranche d'âge spécifiée dans le ménage[38].

4. Calcul des paramètres suivants à partir de ce qui précède:

   c_i nombre de ménages dans la catégorie C de la ZD i

   e_i nombre de ménages dans la catégorie E de la ZD i

   c_m, e_m moyennes respectives de l'expression ci-dessus sur toutes les zones de la «population» (c'est-à-dire dans l'échantillon de la phase 1).

5. Calcul d'une «mesure de la taille» M_i de chaque ZD i comme suit

$$M_i=g1*(c_i/c_m)+(1-g1)*(e_i/e_m),$$

où g1 est un paramètre constant déterminé par l'importance relative dans la sélection des zones à attribuer au nombre de ménages contenant ou susceptibles de contenir des enfants travailleurs comparativement à d'autres ménages. Comme expliqué ci-dessus, une telle mesure de la taille peut être utilisée pour la sélection de PPT pour concemtrer l'échantillon de la phase 2 aux zones à fortes concentrations de travail des enfants.

Il apparaît raisonnable de recommander la valeur de g1=0,8[39] pour le cas particulier actuellement examiné. Les valeurs qui sont raisonnablement plus grandes que 0,5 seraient appropriées dans la plupart des situations.

6.4.3. Taille de l'échantillon

Comme souligné ci-dessus, la taille d'échantillon de la phase 2 devrait être modeste comme cela est prochainement indiqué, avec des additions appropriées permettant la non-réponse.

Ménages de la catégorie C	n_c=6 000
Ménages de la catégorie E	n_e=4 000

[37] Ce terme désigne les «enfants travailleurs» comme défini pour l'enquête avec des limites d'âge spécifiées. Pour l'échantillonnage, ces définitions et leurs implications doivent être prises comme telles.

[38] Le symbole «C» est utilise ici pour designer les ménages qui contiennent réellement ou qui ont une forte probabilité de contenir des enfants travailleurs. Ces ménages sont les principaux fournisseurs des cas d'échantillon pour la composante enfant travailleurs de l'enquête. Le symbole «E» indique d'autres ménages qui ne contiennent pas ou qui ont une faible probabilité de contenir des enfants travailleurs déclarés comme tels lors de l'opération de listage. Ces ménages sont essentiellement pour le module main d'œuvre (emploi).

[39] Une telle recommandation est justifiée sur la base du jugement pratique et de l'expérience antérieure. Avec des données réelles, il devrait être possible d'ajuster cette valeur pour obtenir un échantillon ayant les caractéristiques souhaitées.

Taille totale d'échantillon atteinte $n=n_c+n_e=10\ 000$ ménages.

Ces deux catégories de ménages proviendront bien sûr du même échantillon de ZD. La raison d'être de cette recommandation sur la taille d'échantillon est la suivante.

La plupart de ménages de la catégorie C contiendront un ou plusieurs enfants travailleurs ou potentiellement travailleurs pour que le nombre de tels enfants dans l'enquête soit de l'ordre de 6 000 et peut-être plus. La taille d'échantillon pourra ainsi être de l'ordre de 3 000 enfants dans les zones urbaines et rurales qui sont les principaux domaines de rapport. Avec 4 domaines de rapport, notamment urbains et ruraux selon le sexe ou les principales régions géographiques, on devrait obtenir une taille d'échantillon de l'ordre de 1 500 enfants par domaine. Sur la base de l'expérience, de telles tailles sont suffisantes pour une analyse assez minutieuse.

Pour le module EMO, les deux parties de l'échantillon sont pertinentes. La taille d'échantillon totale disponible avoisinera 10 000 ménages. Etant donné qu'il y a en moyenne 2 personnes économiquement actives par ménage dans le pays, on peut s'attendre à environ 20 000 personnes économiquement actives dans l'échantillon. Cela permettrait la répartition par sexe, par groupe d'âge principal et par zone urbaine/rurale.

Le selection de n_e peut se faire presque indépendamment du choix n_c actuellement abordé ici. Il est question de savoir quel montant du budget et quels efforts on peut consacrer à la composante EMO dans une enquête ciblant principalement le travail des enfants.

6.4.4. Niveaux d'échantillonnage et nombre de grappes

L'échantillon de la phase 2 doit être sélectionné à deux niveaux à partir de la phase 1: la sélection des zones, puis la sélection des ménages. Cela n'exclut pas l'échantillonnage des «grappes compactes» dans certains cas, c'est-à-dire introduire dans l'échantillon tous les ménages de la grappe. Toutes les personnes vivant dans un ménage sélectionné seront incluses dans l'enquête.

Il est très important de se demander: «Combien de grappes sélectionner?». En effet, cette question détermine le coût de l'enquête, la précision de l'échantillonnage et la qualité des données.

Dans le cas de l'EET, il est aussi important de s'assurer qu'un nombre assez grand de grappes est sélectionné pour correspondre à la taille d'échantillon cible des enfants travailleurs. Si l'ampleur du travail des enfants est faible, un grand nombre de grappes devra être sélectionné à cette fin. Toutefois, dans le pays, le travail des enfants semble être très répandu, à telle enseigne qu'un nombre «modeste» de grappes peut suffire.

La procédure de détermination du nombre de grappes, a, devant être sélectionné pour correspondre à la taille d'échantillon n_c est comme suit.

1. Nous devons considérer que la sélection des ZD de la phase 1 à la phase 2 se fera avec la PPT, probabilité proportionnelle à la mesure de la taille M_i définie précédemment.

2. Supposons d'abord que dans chaque zone sélectionnée, tous les ménages de la catégorie C sont introduits dans l'échantillon. A l'aide de cette procédure, nous obtenons le nombre minimal possible de grappes, $a_{min,}$ qui peut correspondre à la taille d'échantillon cible. Cette valeur minimale dépend du nombre moyen de ménages de la catégorie C par ZD dans l'échantillon $c_{m(s)}$.

   $$a_{min}=n_c/c_{m(s)}$$

3. Cette moyenne par grappe dans l'échantillon (de la phase 2) n'est pas la même que la moyenne connue c_m par grappe dans la population (c'est-à-dire dans l'échantillon de la phase 1). La première est plus grande que la dernière, dépendant de la variabilité des mesures de la taille M_i utilisée pour sélectionner les zones.

   Etant donné que l'échantillon de la phase 2 n'a pas été sélectionné à ce niveau, il peut être utilisé pour calculer la moyenne $c_{m(s)}$. Cependant, on peut calculer son *espérance mathématique* à partir de l'information contenue dans la base de sondage comme

   $$c_{m(s)} =\Sigma M_i \cdot c_i /\Sigma M_i$$

   où la somme englobe toutes les grappes de la base de sondage (c'est-à-dire l'échantillon de la phase 1).

   Remarquez que la moyenne de toutes les grappes de la population est, par contre, la moyenne non pondérée

   $$cm=\Sigma c_i/\Sigma 1$$

4. Certaines grappes peuvent contenir trop de ménages ayant des enfants travailleurs et il est généralement conseillé de fixer une limite, soit c_{max}, sur le nombre de ménages de la catégorie C qui seront sélectionnés dans toute ZD. Le fait de fixer une telle limite supérieure augmentera généralement le nombre de grappes, a, nécessaire pour obtenir la taille d'échantillon cible n_c. Ce nombre est le plus petit quand la limite supérieure relevée ci-dessus équivaut ou excède la plus grande des valeurs c_i au sein de la population.

L'important est que, pour une taille d'échantillon cible donnée n_c, il y ait un lien entre les quantités *a* et c_{max}: elles ne peuvent pas être choisies indépendamment. La relation réelle dépend de la distribution empirique des valeurs c_i dans l'échantillon.

On recommande la procédure suivante. Prenez les différentes valeurs c_{max} et calculez le nombre sous-entendu de grappes *a*. Puis, sur la base des considérations pratiques et de ce que l'on trouve dans la base de sondage réelle, on peut choisir le paramètre *a* et obtenir l'expression c_{max} sous-entendue par cette sélection.

On peut le faire à partir de la base de sondage, sans réellement sélectionner l'échantillon de la phase 2. L'algorithme est le suivant.

A l'aide d'une valeur choisie de c_{max}, soit c'_{max}, calculez pour chaque ZD i.

$$c'_i = \min(c_i, c'_{max})$$

L'espérance mathématique du nombre moyen de ménages de la catégorie C de l'échantillon est

$$c'_{m(s)} = \Sigma M_i \cdot c'_i / \Sigma M_i$$

D'un point de vue pratique, il serait souhaitable d'empêcher que le nombre de grappes dans l'échantillon devienne trop grand. Evidemment, on doit choisir $a = a_{min}$ défini dans (2) ci-dessus.

A l'essai, une valeur comme a=400 pourrait être examinée, si on l'estime suffisante pour produire la taille d'échantillon requise pour les ménages de la catégorie C.

Puis la c_{max} sous-entendu correspondant à la valeur choisie de *a* peut être calculée comme expliqué ci-dessus et appliqué plus tard au stade de sélection des ménages. Dans cette illustration n_c=6 000 et a=400, c_{max} peut se révéler être de l'ordre de 20 à 30 ménages.

6.4.5. Sélection des unités de zones

L'échantillon de la phase 2 est obtenu à partir des zones de la phase 1 en sélectionnant les unités avec des probabilités proportionnelles à la mesure de la taille M_i définie ci-dessus. La procédure d'échantillonnage systématique circulaire standard peut être utilisée à cette fin.

Avec *a* comme le nombre de grappes à sélectionner, l'intervalle d'échantillonnage applicable aux mesures cumulées de taille est

$$I = \Sigma M_i / a$$

où la somme englobe toutes les grappes de la base de sondage. La probabilité de sélection d'une zone est

$$f2_i = a \cdot (M_i / \Sigma M_i) = M_i / I$$

C'est la probabilité de sélection conditionnelle de la phase 1 à la phase 2. Soit $f1_{i,}$ la probabilité de la phase 1 de la zone, la probabilité totale d'une zone apparaissant dans l'échantillon de la phase 2 est

$$f1_i * f2_i$$

6.4.6. Sélection des ménages

Comme expliqué ci-dessus, jusqu'à une certaine limite supérieure c_{max}, tous les ménages de la catégorie C dans la grappe sélectionnée seront inclus dans l'échantillon. La probabilité de sélection d'un ménage de la catégorie C dans une zone ayant $c_i = c_{max}$ est la même que celle de sa ZD.

Dans les grappes où le nombre trouvé de ménages de la catégorie C, $c_{i,}$ excède c_{max}, seul c_{max} sera inclus dans l'échantillon. Une telle sélection de ménage peut se faire en recourant à une procédure d'échantillonnage systématique circulaire standard à probabilité égale. Le dernier taux d'échantillonnage du stade est (c_{max}/c_i).

Pour l'un ou l'autre des cas, on définit $c'_i=\min(c_i,c_{max})$,

La probabilité de sélection générale d'un ménage de la catégorie C dans la zone i est

$$f_i=f1_i*f2_i*(c'_i/c_i)$$

6.4.7. Pondération et estimation

Les procédures de pondération et d'estimation, ainsi que les procédures pratiques d'estimation de la variance sont abordées en détail au chapitre 7. On en relève ici quelques points saillants.

Comme dans tout échantillon, il est nécessaire de s'assurer que pour chaque cas d'échantillon une pondération a été convenablement calculé et enregistré dans les microdonnées. Ces pondérations sont inversement proportionnelles aux probabilités de sélection f_i, ajustées pour la non réponse, le calibrage, etc. comme il se doit. Pour la production de «statistiques primaires» comme les moyennes et les ratios, aucune autre référence n'est requise à la structure de l'échantillon, notamment la strate et la grappe auxquelles appartiennent les unités finales.

Les procédures pratiques sont disponibles pour l'estimation de la variance dans les enquêtes dont la taille et le type sont actuellement abordés. Pour l'application de ces procédures, les informations essentielles de la structure de l'échantillon qui doivent être disponibles et théoriquement codées dans les microdonnées, incluent la précision des points suivants pour chaque unité finale: (i) son domaine d'estimation; (ii) sa strate d'échantillonnage; (iii) UEP; (iv) ordre de sélection des UEP, si elles sont systématiquement sélectionnées; et (v) poids de l'échantillon.

Un objectif majeur de l'enquête en phase 2 consiste à améliorer les estimations de la phase 1 sur l'ampleur du travail des enfants en fournissant des facteurs de correction pour sa sous-estimation probable.

Les résultats de la phase 1 sont sujets à l'erreur de «faux positifs» (non déclaration du travail des enfants quand il y en a aucun dans le ménage), mais surtout l'erreur de «faux négatif» (erreur de déclaration du travail des enfants existant). La phase 2 peut être utilisée pour estimer ces taux d'erreur distinctement pour les différentes classes d'enfants et éventuellement aussi pour quelques principaux domaines. Ces facteurs peuvent alors être appliqués aux résultats de la phase 1 pour produire des estimations améliorées. Pour les domaines de rapport plus petits, les facteurs d'ajustement estimés pour le domaine principal auquel ils appartiennent peuvent être utilisés.

6.5. Echantillonnage pour le module EMO

Ce module est simplement une partie d'un entretien intégré et est appliqué dans les mêmes zones comme ci-dessus.

Pour satisfaire les exigences de la taille d'échantillon du module dans l'enquête, d'autres ménages sont sélectionnés dans la catégorie E définie précédemment (section 6.2.2)[40].

6.5.1. Taille d'échantillon

Le principal paramètre à choisir est la taille d'échantillon n_e de la catégorie des ménages E. Il est essentiellement déterminé par les exigences du module EMO. En principe, cette selection peut se faire indépendamment de la sélection de n_c. Normalement, la plupart des ZD contiendra un nombre suffisant de ménages de la catégorie E pour produire le nombre d'échantillon requis de cas.

Cependant, il est souhaitable d'être modéré dans la sélection de cette taille d'échantillon. Remarquez que pour le module EMO, les deux parties de l'échantillon sont utilisées, donnant un total de (n_c+n_e) ménages, et presque deux fois le nombre de personnes économiquement actives.

6.5.2 Probabilités de sélection des ménages

Comme détaillé dans la section 6.4, les zones d'échantillon ont été sélectionnées avec des probabilités $f2_i$ proportionnelles à la mesure de la taille M_i qui est lourdement pondérée en faveur des zones à forte ampleur de travail des enfants (comme reflété par c_i). Cette procédure était appropriée pour les ménages de la catégorie C d'où proviennent les entretiens de l'EET.

Ce sur-échantillonnage des zones à plus forte concentration du travail des enfants est conservé au niveau des ménages de la catégorie C en utilisant l'échantillonnage à tirage complet dans toute la grappe avec $c_i=c_{max}$ défini précédemment. (Section 6.4).

Un tel sur-échantillonnage de ménages avec des hauts niveaux de travail des enfants n'est cependant pas efficace pour la composante EMO. Il est enlevé en sélectionnant les ménages au sein des grappes ayant des probabilités inversement proportionnelles à $f2_i$, annulant ainsi le sur-échantillonnage au stade précédent.

Soit un taux d'échantillonnage

$$f3_i=g2/f2_i$$

[40] Les deux parties de l'échantillon, dans la catégorie C et E respectivement de chaque zone, sont sélectionnées séparément et éventuellement en utilisant différentes procédures, ce n'est qu'alors qu'elles peuvent être assemblées pour former un seul échantillon et être traitées uniformément au stade de mise en œuvre. Les deux parties sont sélectionnées à différents taux au sein de chaque agrégat d'échantillon, étant donné qu'elles sont censées différer grandement dans leurs ampleurs du travail des enfants.

appliqué dans la sélection des ménages de la catégorie E au sein des zones d'échantillon de la phase 2. g2 constant est déterminé par les exigences de la taille d'échantillon. L'espérance mathématique de la taille d'échantillon de la catégorie E peut être estimée avant de sélectionner réellement les zones d'échantillon de la phase 2 comme suit.

La taille totale d'échantillon escomptée

$$n_e=\Sigma e_i \cdot f2_i \cdot f3_i=\Sigma e_i \cdot f2_i \cdot (g2/f2_i)=g2 \cdot \Sigma e_i$$

où la somme englobe toutes les zones de la base de sondage (échantillon de la phase 1). L'expression ci-dessus donne la constante requise g2 comme

$$g2=n_e/\Sigma e_i$$

La probabilité de sélection finale d'un ménage de la catégorie E dans la zone i est

$$f_i=f1_i \cdot f2_i \cdot f3_i=g2 \cdot f1_i$$

Autrement dit, les probabilités de sélection uniformes des zones au sein des domaines de l'échantillon de la phase 1 sont rétablies, en termes relatifs, dans l'échantillon des ménages de la catégorie E.

6.5.3. Réglage de précision

Il est toujours souhaitable pour des raisons pratiques de fixer des limites inférieures et supérieures sur le nombre de ménages de la catégorie C à sélectionner dans toute ZD.

Une limite inférieure e_{min}, peut être considérée souhaitable pour éviter de devoir aller dans une zone pour seulement un ou deux entretiens. Cela se fait en prenant le nombre de ménages à sélectionner comme

$$t_i=\max\,(e_{min}, e_i \cdot f3_i)$$

ou plus précisément comme

$$t_i=\min\,[\max\,(e_{min}, e_i \cdot f3_i), e_i]$$

Similairement, on peut fixer une limite e_{max} sur le nombre maximal de ménages de la catégorie E à introduire dans l'échantillon de toute zone.

Remarquez que l'impact escompté de tels ajustements peut être estimé à partir de la base de sondage avant de sélectionner l'échantillon de la phase 2 en examinant la distribution des valeurs e_i dans la base de sondage après la pondération de leur distribution par M_i.

Evidemment, les probabilités de sélection du niveau final en présence de tels ajustements deviennent

$$f3_i=t_i/e_i$$

La sélection appropriée des limites e_{min} et e_{max} est une question de jugement pratique sous l'éclairage des informations empiriques concernant l'importance de l'existence d'un tel ajustement lorsque cela est nécessaire.

Les valeurs raisonnables dans le cas présent peuvent être une expression du genre

$e_{min}=3$, $e_{max}=20$ à 30

6.6. Illustration numérique des procédures d'échantillonnage

Considérons un domaine hypothétique dans le pays pour lequel des estimations distinctes sont requises sur diverses caractéristiques du travail des enfants. Le domaine se compose de 100 000 ménages dans 1 000 ZD de recensement, chacune d'elle ayant exactement 100 ménages. En réalité, les ZD varient invariablement en taille, mais l'hypothèse ci-dessus est prise dans cette illustration par souci de simplicité.

La phase 1 de l'opération d'enquête se compose de la sélection d'un échantillon de ZD et du listage complet des ménages dans chaque ZD sélectionnée et également l'identification des ménages qui:

- contiennent un ou plusieurs enfants dans le groupe d'âge spécifié engagé dans le travail des enfants, ou.
- ne contiennent aucun enfant déclaré travailleur, mais en contiennent un ou plusieurs qui ne vont pas à l'école.

Ce sont nos ménages de la catégorie C. Le reste dans la ZD a été qualifié de ménages de la catégorie E.

L'échantillon national total de la phase 1 a été ventilé entre les domaines d'une manière appropriée, et nous supposerons que notre domaine reçoit une taille d'échantillon de 100 ZD, c'est-à-dire que 10 pour cent de son total, contenant 10 000 ménages à lister avec un «mini entretien» pour identifier les ménages de la catégorie C.

6.6.1. Sélection des ZD pour la phase 1

Le processus implique la sélection de 100 ZD sur 1 000 dans le domaine avec une probabilité égale, c'est-à-dire avec une constante

$f1=0,1$

Les ZD de la base de sondage sont disposées par domaines urbains et ruraux, et les ZD nomades forment un sous-domaine distinct. Les localités urbaines sont disposées en fonction de la taille et les villages en trois groupes de presqu'une même taille suivant le niveau moyen d'analphabétisme qui est censé être disponible à partir du recensement de la population. Par ailleurs, les ZD sont disposées en fonction de la situation géographique au sein des groupes ci-dessus définis.

La sélection des ZD doit se faire en recourant à l'«échantillonnage systématique circulaire avec une probabilité constante». Si l'on revient aux trois sous-domaines – à savoir urbain, rural et nomade – on se retrouve avec deux options légèrement différentes.

1. La première option consiste à affecter à chaque sous-domaine une taille d'échantillon proportionnelle au nombre de ZD en son sein, arrondie à l'entier le plus proche et sélectionner le nombre ainsi affecté indépendamment au sein de chaque sous domaine. Cela peut occasionner de légères différences de taux d'échantillonnage f1 parmi les domaines en raison de l'arrondissement des nombres à sélectionner. Par exemple, à supposer que le sous-domaine urbain contient 233 ZD, ce qui donne 23,3 comme répartition d'échantillon. Ce chiffre est arrondi à 23 qui devrait être sélectionné. f1 du domaine est donc réduit par le facteur (23/23,3)

2. La deuxième option consiste à placer un sous-domaine après l'autre dans une liste combinée et à sélectionner un échantillon de 100 ZD en une seule opération systématique. Le taux d'échantillonnage f1 reste généralement uniforme, mais le nombre réel sélectionné dans un sous domaine particulier peut varier légèrement par hasard.

La procédure réelle utilisée devrait être enregistrée comme elle est pertinente dans les procédures de calcul de la variance (voir chapitre 7). Plus important encore, il faut tenir un registre de l'ordre de sélection des ZD quand l'échantillonnage systématique a été utilisé.

Les deux options ont leurs mérites, mais admettons ici que par souci de simplicité la procédure (2) a été suivie. La procédure de sélection implique les deux étapes suivantes:

1. Les ZD ordonnées dans la liste comme expliqué ci-dessus sont dénombrées par séquence, dans ce cas de 1 à 1 000.

2. Un nombre aléatoire à trois chiffres est choisi dans la fourchette de 000 à 999. Admettons que ce nombre est 086. Puis que la ZD qui a pour nombre de séquence R=86 est la première à être sélectionnée.

3. L'intervalle d'échantillonnage I est calculé comme le ratio entre le nombre de ZD de la base de sondage et le nombre à sélectionner, soit dans ce cas I=1 000/100=10. Ici cet intervalle se révèle être un nombre entier, ce qui n'est cependant pas toujours le cas -voir (vii) ci-dessus.

4. La suivante ZD sélectionnée est celle dont le nombre de séquence est égal à R+I=96, et celle d'après comme 96+10=106, etc.

5. Lorsque le nombre à sélectionner dépasse 1 000 pour la première fois, réduisez par 1 000 et utilisez le résultat pour sélectionner l'unité suivante, en commençant par

le début de la liste. Dans cet exemple, quand on atteint le nombre 1 006, l'unité 6 est sélectionnée, puis le nombre 16, etc.[41].

6. Poursuivez le processus jusqu'à ce que le nombre requis (100 ici) d'unités soit sélectionné.

7. L'intervalle de sélection I et, bien sûr, le nombre aléatoire initial R doivent être des entiers. Si aucun des deux est non-intégrale, utilisez exactement la même procédure ci-dessus pour construire la séquence R, R+I, R+2·I, …, et arrondissez chaque fois le résultat pour préciser le nombre de séquences de l'unité à sélectionner par la suite.

6.6.2. La base de sondage pour l'échantillonnage de la phase 2

L'échantillonnage de la phase 1 forme la «population» de la phase 2. Les résultats de la phase 1 fournissent la base de sondage correspondante pour la sélection de l'échantillon de la phase 2. La base de sondage est hiérarchisée, avec des données à deux niveaux:

1. Une liste de toutes les ZD de l'échantillon de la phase 1 (assortie de cartes et de descriptions, etc.) avec des information concernant chaque ZD i sur c_i le nombre de ménages de la catégorie C en son sein, et e_i le nombre de ménages restants de la catégorie E.

2. Pour chaque ZD, un listage de tous les ménages s'y trouvant (assorti de leurs adresses, situation géographique, description, etc.), avec des informations sur chaque ménage à savoir s'ils appartiennent à la catégorie C ou à la catégorie E.

En fait, dès que l'échantillon des ZD de la phase 2 a été sélectionné, les listes susmentionnées de ménages ne sont utilisées que pour les zones sélectionnées afin de tirer les échantillons de ménages, en ce qui concerne l'enquête EET-EMO. Toutefois, la liste complète peut être utile pour d'autres enquêtes.

Le tableau 6.7 qui figure plus loin dans ce chapitre fournit un exemple de la base de sondage de la phase 2 au niveau de la ZD, synthétiquement construite pour la présente illustration.

Il contient 100 ZD, numérotées I=1-100. A chaque ZD, les valeurs supposées des variables c_i et e_i définies ci-dessus[42] sont attribuées. Les ZD ont été ordonnées en fonction des valeurs c_i – ce qui est en fait le bon ordre pour la sélection des unités dont la probabilité dépend de c_i.

[41] Remarquez que si le nombre aléatoire d'origine R est 000, il est traité comme 1 000 pour que la première unité à sélectionner soit celle qui est en fin de liste. Puis continuer à partir du début de la liste – la seconde unité sélectionnée étant R+I=0+10=10, autrement dit le nombre d'unités 10.

[42] Ces valeurs ont été simulées aux fins de cette illustration à l'aide des procédures suivantes.
Les unités ont reçu des valeurs x_i à partir de 0 jusqu'à 0,99 par tranches de 0,01, dans leur ordre d'énumération. Les valeurs requises sont calculées comme: $c_i=100\cdot x_i^2$, $e_i=100-c_i$.
Avec cette distribution, la moyenne c_m est un peu en deçà de 100/3. (en fait, il y a 32,8% de ménages ayant des enfants travailleurs ou qui ne vont pas à l'école dans le présent exemple.)

Chaque ZD reçoit une mesure de la taille définie comme

$$M_i=0,8(c_i/c_m)+0,2(e_i/e_m).$$

Manifestement, la valeur moyenne de M_i ainsi définie est 1,0. Les unités seront sélectionnées avec des probabilités proportionnelles à cette mesure de la taille.

6.6.3. Nombre maximal et minimal de ZD sélectionnées

D'ordinaire, un plan à deux degrés est approprié dans le type de situation abordé: sélection d'un sous-échantillon de ZD, puis sélection de ménages au sein de chaque ZD sélectionnée.

Toutefois, le tableau 6.7 illustre aussi les deux schémas de sondage «extrêmes», impliquant chacune une sélection à un seul stade.

Echantillonnage à tirage complet

Ce terme renvoie à la sélection d'un échantillonnage de ZD, mais alors conservant tous les ménages de la catégorie C dans l'échantillon. En clair, pour atteindre la cible de la taille d'échantillon $n_c=600$ ménages de la catégorie C dans notre exemple, le nombre requis de ZD à sélectionner (avec les probabilités proportionnelles spécifiées par rapport à M_i) est le minimum possible, étant donné que chaque ZD sélectionnée apporte la contribution maximale possible (tous ses ménages de la catégorie C) à l'échantillon. Dans le tableau 6.7, la colonne [6] indique le nombre de tels ménages contribués par une ZD *s'il est sélectionné*. Elle est tout simplement égale à la colonne [3], la valeur c_i de la ZD.

Pour apprécier les caractéristiques des échantillons résultant d'un tel plan, il n'est pas nécessaire de sélectionner réellement différents échantillons. Nous pouvons examiner l'apport de chaque ZD à l'échantillon sur la moyenne (c'est-à-dire sur tous les échantillons éventuels pouvant être tirés). *L'espérance mathématique* pour une ZD est donnée par sa valeur réelle (c_i) multipliée par sa probabilité de sélection (M_i)[43]. Elle est indiquée dans la colonne [7]. La moyenne de ces espérances mathématiques est la moyenne $c_{m(s)}$ que nous attendons dans un échantillon. Cela est égale 51,5 ménages de la catégorie C dans notre illustration, qui est considérablement plus grand que la moyenne de toutes les 100 ZD de la population (32,8). Le nombre minimal de ZD à sélectionner qui est censé produire un échantillon de 600 ménages de la catégorie C (si de tels ménages sont prélevés dans chaque zone sélectionnée) est donc

$$a=n_c/c_{m(s)}=600/51,5=11,7$$

Il suffirait de sélectionner 12 ZD et de prélever tous les ménages de la catégorie C dans les ZD sélectionnées dans l'échantillon.

Cette procédure d'identification des propriétés de l'échantillon à partir de la base de sondage, sans sélectionner réellement des échantillons particuliers, est très appropriée pour calculer les conséquences ou estimer les paramètres de diverses options du plan d'échantillonnage. Dans un échantillon, les apports individuels des unités, tels que ceux

[43] En fait, cette probabilité est égale à c_iM_i multiplié par une constante, $a/\Sigma M_i=a/100$ dans notre cas.

fournis dans la colonne [6] sont réalisés pourvu que l'unité concernée vienne réellement à être sélectionnée. Par conséquent, leurs moyennes ou total etc. sont calculés en additionnant les unités réellement sélectionnées dans un *échantillon particulier*. En revanche, toutes les unités au sein de la population contribuent aux quantités du type montré dans la colonne [7]. Ces quantités sont déterminées par les valeurs unitaires (telle que c_i) et la probabilité (M_i) de leur apparition dans tout échantillon. Par conséquent, les moyennes ou les sommes etc. sur toutes les unités de la population indiquent ce que ces statistiques seraient pour un échantillon sur la moyenne qui est sous un plan d'échantillon donné.

Dans l'exemple ci-dessus, l'échantillonnage par tirage complet s'applique aux ménages de la catégorie C. Les ménages de la catégorie E sont plus nombreux et nécessitent généralement le sous-échantillonnage – Voir Sections 6.6.5 et 6.6.6

Conservation de toutes les ZD

C'est ici l'autre extrême: maximiser le nombre de ZD dans l'échantillon et minimiser le nombre de ménages prélevés par ZD. Pour obtenir 600 ménages de la catégorie C à partir de 100 ZD, il nous faut une moyenne de 6,0 par ZD sélectionnée dans l'échantillon. Le point de raccourci $c_{max,}$ c'est-à-dire le nombre minimal de ménages prélevés dans chaque ZD doit être un peu plus élevé que 6,0 parce que certaines ZD ont bien moins que ce nombre de ménages de la catégorie C pour contribuer et les autres ZD doivent le compenser.

Pour avoir le point de raccourci requis, nous pouvons travailler avec la contribution *attendue* de chaque ZD dans le sens entendu ci-dessus, mais bien sûr sous réserve maintenant de la contrainte qu'il ne peut pas dépasser la c_{max} imposée dans chaque ZD de l'échantillon. Une ZD ne peut contribuer que la plus petite de sa valeur réelle c_i et la limite c_{max}, c'est-à-dire que

$$c'_i = \min(c_i, c_{max})$$

L'espérance mathématique, étant donné qu'elle est comme avant $M_i \cdot c'_i$ qui dépend maintenant de la valeur c_{max} choisie. La moyenne de ces valeurs doit être 6,0. A partir de $c_{max}=6{,}0$ environ, on peut ajuster c_{max} de façon itérative en montant jusqu'à ce que cette moyenne soit atteinte[44].

Dans notre illustration, cette valeur se révèle être 6,4. La colonne [8] présente le nombre réel apporté par chaque ZD quand il vient à apparaître dans l'échantillon, avec le raccourci estimé applicable.

[44] Bien sûr que, dans la pratique, on peut travailler avec des nombre entiers. Dans chaque cas, un nombre tel que 6,4 peut être imposé dans la moyenne, par exemple en le considérant comme 7 dans une proportion de 0,4 de tels cas, et comme 6 dans la proportion restante 1-0,4=0,6 proportion sur les cas. Il y a un exemple dans la section 6.7 d'une procédure d'arrondissement d'un ensemble de petits nombres, mais de telle sorte que le cumul d'erreur dans le total résultant soit évité.

6.6.4. Echantillonnage à deux degrés de la phase 1 à la phase 2

Le tableau 6.8 présente des exemples de sélection d'échantillon à deux degrés: un échantillon de ZD est d'abord sélectionné, puis un échantillon de ménages au sein de chaque ZD.

On suppose que dans chaque ZD sélectionnée, tous les ménages de la catégorie C sont conservés dans l'échantillon jusqu'à une certaine limite c_{max}. Si le nombre de ménages disponibles de la catégorie C au sein de la zone (c_i) excède cette limite, seuls les c_{max} sont conservés dans l'échantillon.

Dans le premier exemple du tableau 6.8, on prend c_{max}=30 tel quel, et on calcule le nombre de ZD qui doit être sélectionné pour avoir, sur la moyenne, une taille d'échantillon de n_c ménages de la catégorie C. Ainsi, une zone dans l'échantillon apporte la plus petite des deux valeurs: c_i et c_{max}.

On calcule simplement la contribution attendue de chaque zone de la population (proportionnelle à la colonne [6]). A partir de ces données, le nombre de grappes est calculé directement comme expliqué dans la note de bas de page ci-dessous[45].

Dans le second exemple, nous considérons la situation inverse. On a décidé de sélectionner un certain nombre de grappes (a=35) et l'objectif est de déterminer le maximum de prélèvements par grappe c_{max} pour atteindre la taille d'échantillon cible n_c. Cela nécessite une itération: on ajuste les valeurs de c_{max} jusqu'à ce que le résultat donne a=35, comme requis. Normalement, cela peut se faire facilement sur un tableur. Les résultats sont présentés dans les colonnes [9] et [10] du tableau 6.8. Pour n_c=600 et a=35, la valeur requise c_{max}=19,3. En considérant que c_{max}=20, un échantillon de a=34 ZD suffit.

6.6.5. Echantillonnage des ménages de la catégorie E

Les illustrations supposent le plan suivant: la sélection de ZD est déterminée en premier par les exigences de l'échantillon des ménages de la catégorie C. Dès que les ZD sont sélectionnées, l'échantillonnage des ménages de la catégorie E est exécuté pour atteindre la taille d'échantillon requise n_e. Au sein des zones sélectionnées, le taux d'échantillonnage de cette catégorie de ménages est inversement proportionnel à la mesure de la taille M_i avec laquelle la zone a été sélectionnée dans la phase 1. *Au sein d'une zone*, les ménages sont normalement sélectionnés avec des probabilités égales, en utilisant par exemple l'échantillonnage systématique circulaire à probabilité constante.

Ce plan est appliqué aux deux exemples du tableau 6.8 des colonnes [7]-[8] et des colonnes [11]-[12].

[45] Les valeurs réelles des nombres attendus sont (a/100) fois ceux de cette colonne. Cette forme de présentation est appropriée dans le calcul, étant donné que les chiffres résultants ne dépendent pas de a, le nombre de grappes à sélectionner qu'il faut déterminer. La taille d'échantillon requise n_c divisée par la somme des apports indiqués dans col.[9] donne la valeur requise de (a/100).

Remarquez qu'en raison du plan spécifique adopté, le nombre se trouvant dans la colonne [8] est devenu proportionnel au nombre de ménages E que la ZD apporte sur la moyenne[46].

6.6.6. Fixation des limites sur le nombre de ménages E à sélectionner dans une zone

De telles limites sont souvent requises pour des raisons pratiques. Le tableau 6.9 calcule les paramètres du plan d'échantillonnage dans une situation où on a décidé de prendre au moins $e_{min}=3$ ménages de la catégorie E dans l'échantillon d'une ZD une fois qu'elle a été sélectionnée.

Les colonnes [5]-[8] du tableau 6.9 présentent les résultats d'un plan ayant des paramètres tels que définis dans le tableau ($c_{max}=10$, résultant à un échantillon dont la valeur requise *a* avoisine 65). Ces tableaux sont construits de la même manière que le tableau 6.8. Les colonnes [9]-[12] calculent les résultats lorsque la limite $e_{min}=3$ a été imposée. Cette opération n'altère l'échantillon de grappes en aucun cas comme il a déjà été déterminé par l'échantillon C. La fixation d'une limite inférieure e_{min} signifie que ces zones qui auraient moins contribuées que ce nombre de ménages E, contribuent à présent davantage. Cela augmente la taille d'échantillon totale résultante au-delà du nombre n_e (colonnes [9]-[10]). La taille de l'échantillon peut être réduite à la taille requise par simple diminution proportionnelle générale des taux d'échantillonnage des ménages E C'est ce que montre les colonnes [11]-[12]. Dans notre exemple, l'effet d'imposer e_{min} était assez considérable, nécessitant que le taux d'échantillonnage final des ménages de la catégorie soit réduit par le facteur 0,72.

Le tableau 6.10 présente un exemple semblable quand il y a une limite supérieure e_{max} sur le nombre de ménages E dans toute zone. En fait, le cas illustré dans ce tableau concerne la limite supérieure imposée simplement de façon automatique par la condition selon laquelle le nombre de ménages E à sélectionner dans une zone ne peut excéder le nombre de tels ménages disponibles dans cette zone. La procédure reste la même, indépendamment de la raison particulière de la limite supérieure.

Dans le tableau 6.10, les colonnes [7]-[8] appliquent la procédure habituelle sans considérer la limite e_{max}. Les colonnes [9]-[10] recalcule cette donnée avec la limite imposée. Certains ménages contribuent à présent à bien peu de ménages, ce qui réduit la taille d'échantillon résultante en deçà de la cible (colonnes [9]-[10]). La solution consiste à augmenter le taux d'échantillonnage final (presque) proportionnellement pour obtenir la taille d'échantillon requise. C'est ce que montrent les colonnes [11]-[12].

6.6.7. Résumé des simulations

Le tableau 6.5 présente les principaux résultats des simulations décrites ci-dessus: comment les paramètres résultants de l'échantillon changent selon les diverses

[46] Comme avant, la contribution réelle est (a/100) fois les valeurs présentées. Par conséquent, la somme des valeurs présentées sur toutes les ZD de la population donne la taille d'échantillon requise n_c fois (100/a). Etant donné n_c, le paramétre requis a peut être calculé. La procédure a été utilisée par souci de rendre indépendantes de la valeur particulière de a les valeurs présentées dans col.[8].

contraintes. Ce ne sont que quelques chiffres clés des tableaux détaillés 6.7-6.11 qui se trouvent à la fin de ce chapitre.

Un dernier mot sur les valeurs e_{max}. Chaque plan implique une limite supérieure sur le nombre de ménage E sélectionnés dans toute zone, même quand une telle limite n'est pas explicitement imposée. Cela résulte simplement du fait que dans nos exemples, les zones sont sélectionnées principalement du point de vue des ménages C et tous les ajustements généraux apportés aux taux d'échantillonnage des ménages E doivent se faire au sein des zones déjà sélectionnées pour satisfaire aux exigences[47] de la taille d'échantillon (n_e).

Tableau 6.5: Caractéristiques sommaires des diverses illustrations de la base de sondage

	Nombre de grappes	cmax	emax	Moyenne 'C' M par ZD	Moyenne 'E' M par ZD	Nombre total de 'C' M (n_c)	Nombre total de M 'E' (n_e)
Population	100	98,0	100,0	32,8	67,2	3,284	6,717
Echantillonnage à tirage complet	11,7	98,0	100,0	51,5	25,7	600	300
Toutes les ZD conservées dans l'échantillon	100	6,4	15,0	6,0	3,0	600	300
cmax=30	23,7	30,0	63,2	25,3	12,6	600	300
ZD dans l'échantillon, a=35	35,0	19,3	42,9	17,14	8,6	600	300
cmax=10	64,7	10	23,2	9,3	4,6	600	300
cmax=10, emin=3	64,7	10	16,6	9,3	4,6	600	300

6.7. Illustration de la sélection d'échantillon (échantillonnage systématique circulaire PPT)

Le tableau 6.11 présente la méthode d'«échantillonnage systématique circulaire couramment utilisé avec des probabilités proportionnelles à la taille» pour sélectionner les unités de zones. Nous l'appliquons à notre base de sondage de 100 ZD. Un nombre spécifié *a* de ZD sera sélectionné avec des probabilités proportionnelles à la taille M_i présentée dans la colonne [4] du tableau.

Dans tout échantillonnage systématique, la première étape consiste à disposer les unités dans l'ordre requis pour maximiser les avantages de la stratification implicite fournie par l'échantillonnage systématique. Dans le tableau 6.11, les zones sont disposées suivant la concentration croissante de ménages ayant des enfants travailleurs (variable c_i).

Pour illustrer l'utilisation des intervalles de sélection fractionnels et des points de départ de la sélection systématique, supposons qu'un échantillon de a=22 grappes sur un total de 100 doit être sélectionné. Les mesures de la taille M_i s'élèvent à un total de

[47] Avec des mesures de la taille pour la sélection des ZD tel que supposé dans ces illustrations, à savoir $M_i=0.8\cdot(c_i/c_m)+0.2\cdot(e_i/e_m)$, on peut montrer que e_{max} n'excède pas 5 (=1/0.2) fois c_m, le nombre moyen de ménages E par zone dans la population.

T=100 dans notre illustration, la moyenne par zone étant 1,0. On suppose que les valeurs individuelles M_i sont disponibles et utilisées à 2 décimales.

La colonne [5] indique le cumul T_i des valeurs M_i. Avec $T_1=M_1$, on a $T_2=T_1+M_2$, $T_3=T_2+M_3$, … $T_i=T_{i-1}+M_i$, jusqu'à $T_{100}=T$

Pour entamer la procédure systématique circulaire, sélectionnez un nombre aléatoire à 4 chiffres R dans la fourchette 0000-9999 (il y a un point décimal implicite avant les deux derniers chiffres).

Cela définit le premier «point de sélection» $S_1=R$. Une unité particulière, disons j, est identifiée pour la sélection par ce point de sélection. C'est l'unité pour laquelle les mesures de taille cumulatives T_j remplissent la condition suivante (il ne peut y avoir qu'une et une seule limite de ce genre dans la liste):

$$T_{j-1}<R\leq T_j$$

Dans notre exemple, R=64,91, et le nombre d'unités 84 remplit les conditions ci-dessus[48]. Pour cette unité

$$T_{83}=64{,}74,\ T_{85}=66{,}51$$

Il est très important de conserver un registre de l'*ordre de sélection* des unités dans l'échantillon comme présenté dans la colonne [7]. Pour cette unité, l'ordre de sélection est 1.

L'intervalle d'échantillonnage requis est I=T/a=4,55 dans cet exemple.

Le second point de sélection est donné par

$$S2=R+I=69{,}46$$

Cette expression identifie la prochaine unité à sélectionner comme étant 86 avec une mesure de taille cumulée $T_{86}=70{,}16\geq S_2$, alors que $T_{85}=68{,}32<S_2$.

Similairement, le troisième point de sélection est

$$S3=S2+I=R+2I=74{,}00$$

et cette expression identifie le nombre unitaire 89 pour la sélection.

En général, pour le point de sélection k-th

$$S_k=S_{(k-1)}+I=R+(k-1)\cdot I$$

Quand le point de sélection excède le total cumulatif T pour la première fois, il est défini à nouveau en le réduisant par T: quand ST, redéfinissez S>S-T puis poursuivez le processus.

[48] Un point mineur: si R vient à être «0000», cela implique la sélection de la dernière unité de la liste. Après quoi, on poursuit la sélection à partir du haut de la liste.

Dans notre exemple,

$$S_9=S_8+I=96,73+4,55=101,27$$

redéfinissez $S_9=101,27-100,00=1,27$

qui a identifié le nombre unitaire 5 comme étant la suivante (9e) sélection. Ce processus prend fin après la sélection du nombre unitaire 81, quand on a eu l'échantillon requis de 22 unités.

Par voie de conséquence, l'échantillonnage systématique circulaire avec une probabilité constante n'est qu'un cas particulier de ce qui précède. Dans l'exemple ci-dessus, on peut affecter les nombres de 0001 à 0100 à l'unité 1, de 0101 à 0200 à l'unité 2, etc. et de 9901 à (1)0000 à l'unité 100. Comme dans l'exemple ci-dessus, l'origine choisi au hasard R dans la fourchette 0000-09999 et l'intervalle d'échantillonnage I=10000/22 donnent un échantillon à probabilité égale de 22 unités.

Arrondissement

Dans toute application, le nombre d'unités sélectionné doit être un nombre entier. Cela n'est pas toujours important, mais peut faire la différence lorsque le nombre d'unités impliqué est petit.

Quand de petits nombres fractionnels d'une longue séquence doivent être arrondis, leur total peut être affecté de façon significative si les nombres sont arrondis indépendamment les uns des autres. Le tableau 6.6 indique comment le total peut mieux être contrôlé. Les nombres de la colonne [1] sont d'abord cumulés dans la colonne [2]. Les nombres *cumulés* sont ensuite arrondis dans la colonne [3]. Les nombres arrondis sont décumulés dans la colonne [4] pour donner les résultats. La dé-cumulation veut tout simplement dire inverser le processus de cumul

$$T_i=T_{i-1}+M_i$$

en

$$M_i=T_i-T_{i-1}$$

Tableau 6.6: Exemple de procédure d'arrondissement d'une séquence de petits nombres

	(1)	(2)	(3)	(4)
	Nombres fractionnels	Cumul	Cumul arrondi	'Décumul'
	0,3	0,3	0,0	0,0
	0,4	0,6	1,0	1,0
	0,5	1,1	1,0	0,0
	0,6	1,7	2,0	1,0
	0,8	2,6	3,0	1,0
	1,0	3,6	4,0	1,0
	1,2	4,8	5,0	1,0
	1,4	6,2	6,0	1,0
	1,7	7,9	8,0	2,0
	2,0	9,9	10,0	2,0
	2,3	12,1	12,0	2,0
	2,6	14,7	15,0	3,0
	2,9	17,6	18,0	3,0
	3,2	20,8	21,0	3,0
	3,6	24,4	24,0	3,0
	4,0	28,4	28,0	4,0
	4,4	32,8	33,0	5,0
	4,8	37,7	38,0	5,0
	5,3	42,9	43,0	5,0
	5,8	48,7	49,0	6,0
Total	48,7			49,0

Le décumul fournit une série de valeurs arrondies de nombres tels que leur total est presque conservé.

Tableau 6.7: Exemple type de base de sondage des ZD pour l'échantillonnage de la phase 2

Base de sondage					Deux bases de sondage extrêmes			
Paramètres g1= 0,8 nc= 600					Echantillonnage à tirage complet Aucune limite maximale cmax		Toutes les ZD conservées dans l'échantillon cmax= 6,4	
					Nombre de 'C' m sélectionné:		Nombre de 'C' m sélectionné:	
					Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
1	0,00	0,0	100,0	0,30	0,0	0,0	0,0	0,0
2	0,01	0,0	100,0	0,30	0,0	0,0	0,0	0,0
3	0,02	0,0	100,0	0,30	0,0	0,0	0,0	0,0
4	0,03	0,1	99,9	0,30	0,1	0,0	0,1	0,0
5	0,04	0,2	99,8	0,30	0,2	0,0	0,2	0,0
6	0,05	0,3	99,8	0,30	0,3	0,1	0,3	0,1
7	0,06	0,4	99,6	0,31	0,4	0,1	0,4	0,1
8	0,07	0,5	99,5	0,31	0,5	0,2	0,5	0,2
9	0,08	0,6	99,4	0,31	0,6	0,2	0,6	0,2
10	0,09	0,8	99,2	0,32	0,8	0,3	0,8	0,3
11	0,10	1,0	99,0	0,32	1,0	0,3	1,0	0,3

Base de sondage Paramètres g1= 0,8 nc= 600					Deux bases de sondage extrêmes			
					Echantillonnage à tirage complet Aucune limite maximale cmax Nombre de 'C' m sélectionné:		Toutes les ZD conservées dans l'échantillon cmax= 6,4 Nombre de 'C' m sélectionné:	
					Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
12	0,11	1,2	98,8	0,32	1,2	0,4	1,2	0,4
13	0,12	1,4	98,6	0,33	1,4	0,5	1,4	0,5
14	0,13	1,7	98,3	0,33	1,7	0,6	1,7	0,6
15	0,14	2,0	98,0	0,34	2,0	0,7	2,0	0,7
16	0,15	2,3	97,8	0,35	2,3	0,8	2,3	0,8
17	0,16	2,6	97,4	0,35	2,6	0,9	2,6	0,9
18	0,17	2,9	97,1	0,36	2,9	1,0	2,9	1,0
19	0,18	3,2	96,8	0,37	3,2	1,2	3,2	1,2
20	0,19	3,6	96,4	0,37	3,6	1,4	3,6	1,4
21	0,20	4,0	96,0	0,38	4,0	1,5	4,0	1,5
22	0,21	4,4	95,6	0,39	4,4	1,7	4,4	1,7
23	0,22	4,8	95,2	0,40	4,8	1,9	4,8	1,9
24	0,23	5,3	94,7	0,41	5,3	2,2	5,3	2,2
25	0,24	5,8	94,2	0,42	5,8	2,4	5,8	2,4
26	0,25	6,3	93,8	0,43	6,3	2,7	6,3	2,7
27	0,26	6,8	93,2	0,44	6,8	3,0	6,4	2,8
28	0,27	7,3	92,7	0,45	7,3	3,3	6,4	2,9
29	0,28	7,8	92,2	0,47	7,8	3,6	6,4	3,0
30	0,29	8,4	91,6	0,48	8,4	4,0	6,4	3,0
31	0,30	9,0	91,0	0,49	9,0	4,4	6,4	3,1
32	0,31	9,6	90,4	0,50	9,6	4,8	6,4	3,2
33	0,32	10,2	89,8	0,52	10,2	5,3	6,4	3,3
34	0,33	10,9	89,1	0,53	10,9	5,8	6,4	3,4
35	0,34	11,6	88,4	0,55	11,6	6,3	6,4	3,5
36	0,35	12,3	87,8	0,56	12,3	6,9	6,4	3,6
37	0,36	13,0	87,0	0,57	13,0	7,5	6,4	3,7
38	0,37	13,7	86,3	0,59	13,7	8,1	6,4	3,8
39	0,38	14,4	85,6	0,61	14,4	8,8	6,4	3,9
40	0,39	15,2	84,8	0,62	15,2	9,5	6,4	4,0
41	0,40	16,0	84,0	0,64	16,0	10,2	6,4	4,1
42	0,41	16,8	83,2	0,66	16,8	11,0	6,4	4,2
43	0,42	17,6	82,4	0,68	17,6	11,9	6,4	4,3
44	0,43	18,5	81,5	0,69	18,5	12,8	6,4	4,4
45	0,44	19,4	80,6	0,71	19,4	13,8	6,4	4,5
46	0,45	20,3	79,8	0,73	20,3	14,8	6,4	4,6
47	0,46	21,2	78,8	0,75	21,2	15,9	6,4	4,8
48	0,47	22,1	77,9	0,77	22,1	17,0	6,4	4,9

Base de sondage Paramètres g1= 0,8 nc= 600					Deux bases de sondage extrêmes			
					Echantillonnage à tirage complet Aucune limite maximale cmax Nombre de 'C' m sélectionné:		Toutes les ZD conservées dans l'échantillon cmax= 6,4 Nombre de 'C' m sélectionné:	
					Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
49	0,48	23,0	77,0	0,79	23,0	18,2	6,4	5,0
50	0,49	24,0	76,0	0,81	24,0	19,5	6,4	5,2
51	0,50	25,0	75,0	0,83	25,0	20,8	6,4	5,3
52	0,51	26,0	74,0	0,85	26,0	22,2	6,4	5,4
53	0,52	27,0	73,0	0,88	27,0	23,7	6,4	5,6
54	0,53	28,1	71,9	0,90	28,1	25,2	6,4	5,7
55	0,54	29,2	70,8	0,92	29,2	26,9	6,4	5,9
56	0,55	30,3	69,8	0,94	30,3	28,6	6,4	6,0
57	0,56	31,4	68,6	0,97	31,4	30,4	6,4	6,2
58	0,57	32,5	67,5	0,99	32,5	32,3	6,4	6,3
59	0,58	33,6	66,4	1,02	33,6	34,2	6,4	6,5
60	0,59	34,8	65,2	1,04	34,8	36,3	6,4	6,6
61	0,60	36,0	64,0	1,07	36,0	38,4	6,4	6,8
62	0,61	37,2	62,8	1,09	37,2	40,7	6,4	7,0
63	0,62	38,4	61,6	1,12	38,4	43,0	6,4	7,1
64	0,63	39,7	60,3	1,15	39,7	45,5	6,4	7,3
65	0,64	41,0	59,0	1,17	41,0	48,1	6,4	7,5
66	0,65	42,3	57,8	1,20	42,3	50,8	6,4	7,6
67	0,66	43,6	56,4	1,23	43,6	53,6	6,4	7,8
68	0,67	44,9	55,1	1,26	44,9	56,5	6,4	8,0
69	0,68	46,2	53,8	1,29	46,2	59,5	6,4	8,2
70	0,69	47,6	52,4	1,32	47,6	62,7	6,4	8,4
71	0,70	49,0	51,0	1,35	49,0	65,9	6,4	8,6
72	0,71	50,4	49,6	1,38	50,4	69,4	6,4	8,7
73	0,72	51,8	48,2	1,41	51,8	72,9	6,4	8,9
74	0,73	53,3	46,7	1,44	53,3	76,6	6,4	9,1
75	0,74	54,8	45,2	1,47	54,8	80,4	6,4	9,3
76	0,75	56,3	43,8	1,50	56,3	84,4	6,4	9,5
77	0,76	57,8	42,2	1,53	57,8	88,5	6,4	9,7
78	0,77	59,3	40,7	1,57	59,3	92,8	6,4	10,0
79	0,78	60,8	39,2	1,60	60,8	97,3	6,4	10,2
80	0,79	62,4	37,6	1,63	62,4	101,9	6,4	10,4
81	0,80	64,0	36,0	1,67	64,0	106,7	6,4	10,6
82	0,81	65,6	34,4	1,70	65,6	111,6	6,4	10,8
83	0,82	67,2	32,8	1,74	67,2	116,7	6,4	11,0
84	0,83	68,9	31,1	1,77	68,9	122,0	6,4	11,3
85	0,84	70,6	29,4	1,81	70,6	127,5	6,4	11,5
86	0,85	72,3	27,8	1,84	72,3	133,2	6,4	11,7

Base de sondage Paramètres g1= 0,8 nc= 600					Deux bases de sondage extrêmes			
					Echantillonnage à tirage complet Aucune limite maximale cmax Nombre de 'C' m sélectionné:		Toutes les ZD conservées dans l'échantillon cmax= 6,4 Nombre de 'C' m sélectionné:	
					Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
i	xi	ci=100*x2	ei=100-ci	Mi	ci'	ci' * Mi	ci'	ci' * Mi
87	0,86	74,0	26,0	1,88	74,0	139,0	6,4	11,9
88	0,87	75,7	24,3	1,92	75,7	145,1	6,4	12,2
89	0,88	77,4	22,6	1,95	77,4	151,3	6,4	12,4
90	0,89	79,2	20,8	1,99	79,2	157,8	6,4	12,7
91	0,90	81,0	19,0	2,03	81,0	164,4	6,4	12,9
92	0,91	82,8	17,2	2,07	82,8	171,3	6,4	13,2
93	0,92	84,6	15,4	2,11	84,6	178,4	6,4	13,4
94	0,93	86,5	13,5	2,15	86,5	185,7	6,4	13,7
95	0,94	88,4	11,6	2,19	88,4	193,3	6,4	13,9
96	0,95	90,3	9,8	2,23	90,3	201,1	6,4	14,2
97	0,96	92,2	7,8	2,27	92,2	209,1	6,4	14,4
98	0,97	94,1	5,9	2,31	94,1	217,4	6,4	14,7
99	0,98	96,0	4,0	2,35	96,0	225,9	6,4	14,9
100	0,99	98,0	2,0	2,39	98,0	234,6	6,4	15,2
	Moyenne 'G' M par ZD							
		32,8	67,2	1,00	32,8	51,5		6,0
	Nombre de ZD sélectionnées					11,7		100,0
	Taille d'échantillon attendu (nc)					600,0		600,0

Tableau 6.8: Exemple d'échantillonnage à deux degrés de la phase 1 à la phase 2

Paramètres d'échantillonnage nc= 600 ne= 300				Spécifié: cmax= 30,0 Résultat: ZD dans l'échantillon, a= 23,7				Spécifié: ZD dans l'échantillon, a= 35,0 Result: cmax= 19,3			
				Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi
1	0,0	100,0	0,3	0,0	0,0	63,2	18,8	0,0	0,0	42,9	12,8
2	0,0	100,0	0,3	0,0	0,0	63,2	18,8	0,0	0,0	42,8	12,8
3	0,0	100,0	0,3	0,0	0,0	63,0	18,8	0,0	0,0	42,7	12,8
4	0,1	99,9	0,3	0,1	0,0	62,8	18,8	0,1	0,0	42,5	12,8
5	0,2	99,8	0,3	0,2	0,0	62,4	18,8	0,2	0,0	42,3	12,7
6	0,3	99,8	0,3	0,3	0,1	61,9	18,8	0,3	0,1	42,0	12,7
7	0,4	99,6	0,3	0,4	0,1	61,4	18,8	0,4	0,1	41,6	12,7
8	0,5	99,5	0,3	0,5	0,2	60,8	18,7	0,5	0,2	41,2	12,7
9	0,6	99,4	0,3	0,6	0,2	60,1	18,7	0,6	0,2	40,7	12,7
10	0,8	99,2	0,3	0,8	0,3	59,3	18,7	0,8	0,3	40,2	12,7
11	1,0	99,0	0,3	1,0	0,3	58,4	18,6	1,0	0,3	39,6	12,6
12	1,2	98,8	0,3	1,2	0,4	57,5	18,6	1,2	0,4	39,0	12,6
13	1,4	98,6	0,3	1,4	0,5	56,5	18,6	1,4	0,5	38,3	12,6
14	1,7	98,3	0,3	1,7	0,6	55,4	18,5	1,7	0,6	37,6	12,5
15	2,0	98,0	0,3	2,0	0,7	54,3	18,5	2,0	0,7	36,8	12,5
16	2,3	97,8	0,3	2,3	0,8	53,2	18,4	2,3	0,8	36,1	12,5
17	2,6	97,4	0,4	2,6	0,9	52,0	18,3	2,6	0,9	35,3	12,4
18	2,9	97,1	0,4	2,9	1,0	50,8	18,3	2,9	1,0	34,5	12,4
19	3,2	96,8	0,4	3,2	1,2	49,6	18,2	3,2	1,2	33,6	12,3
20	3,6	96,4	0,4	3,6	1,4	48,4	18,1	3,6	1,4	32,8	12,3
21	4,0	96,0	0,4	4,0	1,5	47,1	18,1	4,0	1,5	32,0	12,3
22	4,4	95,6	0,4	4,4	1,7	45,9	18,0	4,4	1,7	31,1	12,2
23	4,8	95,2	0,4	4,8	1,9	44,6	17,9	4,8	1,9	30,3	12,1
24	5,3	94,7	0,4	5,3	2,2	43,4	17,8	5,3	2,2	29,4	12,1
25	5,8	94,2	0,4	5,8	2,4	42,1	17,7	5,8	2,4	28,6	12,0
26	6,3	93,8	0,4	6,3	2,7	40,9	17,6	6,3	2,7	27,7	12,0
27	6,8	93,2	0,4	6,8	3,0	39,7	17,6	6,8	3,0	26,9	11,9
28	7,3	92,7	0,5	7,3	3,3	38,5	17,5	7,3	3,3	26,1	11,8
29	7,8	92,2	0,5	7,8	3,6	37,3	17,3	7,8	3,6	25,3	11,8
30	8,4	91,6	0,5	8,4	4,0	36,1	17,2	8,4	4,0	24,5	11,7
31	9,0	91,0	0,5	9,0	4,4	34,9	17,1	9,0	4,4	23,7	11,6
32	9,6	90,4	0,5	9,6	4,8	33,8	17,0	9,6	4,8	22,9	11,5
33	10,2	89,8	0,5	10,2	5,3	32,7	16,9	10,2	5,3	22,2	11,5
34	10,9	89,1	0,5	10,9	5,8	31,6	16,8	10,9	5,8	21,4	11,4
35	11,6	88,4	0,5	11,6	6,3	30,5	16,6	11,6	6,3	20,7	11,3
36	12,3	87,8	0,6	12,3	6,9	29,5	16,5	12,3	6,9	20,0	11,2

Paramètres d'échantillonnage nc= 600 ne= 300				Spécifié: cmax= 30,0 Résultat: ZD dans l'échantillon, a= 23,7				Spécifié: ZD dans l'échantillon, a= 35,0 Result: cmax= 19,3			
				Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi
37	13,0	87,0	0,6	13,0	7,5	28,5	16,4	13,0	7,5	19,3	11,1
38	13,7	86,3	0,6	13,7	8,1	27,5	16,2	13,7	8,1	18,7	11,0
39	14,4	85,6	0,6	14,4	8,8	26,6	16,1	14,4	8,8	18,0	10,9
40	15,2	84,8	0,6	15,2	9,5	25,6	16,0	15,2	9,5	17,4	10,8
41	16,0	84,0	0,6	16,0	10,2	24,7	15,8	16,0	10,2	16,8	10,7
42	16,8	83,2	0,7	16,8	11,0	23,8	15,7	16,8	11,0	16,2	10,6
43	17,6	82,4	0,7	17,6	11,9	23,0	15,5	17,6	11,9	15,6	10,5
44	18,5	81,5	0,7	18,5	12,8	22,1	15,3	18,5	12,8	15,0	10,4
45	19,4	80,6	0,7	19,4	13,8	21,3	15,2	19,3	13,8	14,5	10,3
46	20,3	79,8	0,7	20,3	14,8	20,5	15,0	19,3	14,1	13,9	10,2
47	21,2	78,8	0,8	21,2	15,9	19,8	14,8	19,3	14,5	13,4	10,1
48	22,1	77,9	0,8	22,1	17,0	19,0	14,7	19,3	14,9	12,9	9,9
49	23,0	77,0	0,8	23,0	18,2	18,3	14,5	19,3	15,3	12,4	9,8
50	24,0	76,0	0,8	24,0	19,5	17,6	14,3	19,3	15,7	12,0	9,7
51	25,0	75,0	0,8	25,0	20,8	17,0	14,1	19,3	16,1	11,5	9,6
52	26,0	74,0	0,9	26,0	22,2	16,3	13,9	19,3	16,5	11,1	9,4
53	27,0	73,0	0,9	27,0	23,7	15,7	13,7	19,3	16,9	10,6	9,3
54	28,1	71,9	0,9	28,1	25,2	15,1	13,5	19,3	17,4	10,2	9,2
55	29,2	70,8	0,9	29,2	26,9	14,5	13,3	19,3	17,8	9,8	9,0
56	30,3	69,8	0,9	30,0	28,3	13,9	13,1	19,3	18,3	9,4	8,9
57	31,4	68,6	1,0	30,0	29,1	13,3	12,9	19,3	18,7	9,0	8,8
58	32,5	67,5	1,0	30,0	29,8	12,8	12,7	19,3	19,2	8,7	8,6
59	33,6	66,4	1,0	30,0	30,5	12,3	12,5	19,3	19,7	8,3	8,5
60	34,8	65,2	1,0	30,0	31,3	11,8	12,3	19,3	20,1	8,0	8,3
61	36,0	64,0	1,1	30,0	32,0	11,3	12,0	19,3	20,6	7,6	8,2
62	37,2	62,8	1,1	30,0	32,8	10,8	11,8	19,3	21,1	7,3	8,0
63	38,4	61,6	1,1	30,0	33,6	10,3	11,6	19,3	21,6	7,0	7,9
64	39,7	60,3	1,1	30,0	34,4	9,9	11,4	19,3	22,2	6,7	7,7
65	41,0	59,0	1,2	30,0	35,2	9,5	11,1	19,3	22,7	6,4	7,5
66	42,3	57,8	1,2	30,0	36,0	9,0	10,9	19,3	23,2	6,1	7,4
67	43,6	56,4	1,2	30,0	36,9	8,6	10,6	19,3	23,8	5,9	7,2
68	44,9	55,1	1,3	30,0	37,7	8,2	10,4	19,3	24,3	5,6	7,0
69	46,2	53,8	1,3	30,0	38,6	7,9	10,1	19,3	24,9	5,3	6,9
70	47,6	52,4	1,3	30,0	39,5	7,5	9,9	19,3	25,4	5,1	6,7
71	49,0	51,0	1,3	30,0	40,4	7,1	9,6	19,3	26,0	4,8	6,5
72	50,4	49,6	1,4	30,0	41,3	6,8	9,3	19,3	26,6	4,6	6,3
73	51,8	48,2	1,4	30,0	42,2	6,4	9,1	19,3	27,2	4,4	6,1

Paramètres d'échantillonnage nc= 600 ne= 300				Spécifié: cmax= 30,0 Résultat: ZD dans l'échantillon, a= 23,7				Spécifié: ZD dans l'échantillon, a= 35,0 Result: cmax= 19,3			
				Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi
74	53,3	46,7	1,4	30,0	43,1	6,1	8,8	19,3	27,8	4,1	6,0
75	54,8	45,2	1,5	30,0	44,1	5,8	8,5	19,3	28,4	3,9	5,8
76	56,3	43,8	1,5	30,0	45,0	5,5	8,2	19,3	29,0	3,7	5,6
77	57,8	42,2	1,5	30,0	46,0	5,2	8,0	19,3	29,6	3,5	5,4
78	59,3	40,7	1,6	30,0	47,0	4,9	7,7	19,3	30,3	3,3	5,2
79	60,8	39,2	1,6	30,0	48,0	4,6	7,4	19,3	30,9	3,1	5,0
80	62,4	37,6	1,6	30,0	49,0	4,3	7,1	19,3	31,5	2,9	4,8
81	64,0	36,0	1,7	30,0	50,0	4,1	6,8	19,3	32,2	2,8	4,6
82	65,6	34,4	1,7	30,0	51,0	3,8	6,5	19,3	32,9	2,6	4,4
83	67,2	32,8	1,7	30,0	52,1	3,6	6,2	19,3	33,5	2,4	4,2
84	68,9	31,1	1,8	30,0	53,1	3,3	5,9	19,3	34,2	2,2	4,0
85	70,6	29,4	1,8	30,0	54,2	3,1	5,5	19,3	34,9	2,1	3,8
86	72,3	27,8	1,8	30,0	55,3	2,8	5,2	19,3	35,6	1,9	3,5
87	74,0	26,0	1,9	30,0	56,4	2,6	4,9	19,3	36,3	1,8	3,3
88	75,7	24,3	1,9	30,0	57,5	2,4	4,6	19,3	37,0	1,6	3,1
89	77,4	22,6	2,0	30,0	58,6	2,2	4,2	19,3	37,8	1,5	2,9
90	79,2	20,8	2,0	30,0	59,8	2,0	3,9	19,3	38,5	1,3	2,7
91	81,0	19,0	2,0	30,0	60,9	1,8	3,6	19,3	39,2	1,2	2,4
92	82,8	17,2	2,1	30,0	62,1	1,6	3,2	19,3	40,0	1,1	2,2
93	84,6	15,4	2,1	30,0	63,2	1,4	2,9	19,3	40,7	0,9	2,0
94	86,5	13,5	2,1	30,0	64,4	1,2	2,5	19,3	41,5	0,8	1,7
95	88,4	11,6	2,2	30,0	65,6	1,0	2,2	19,3	42,3	0,7	1,5
96	90,3	9,8	2,2	30,0	66,8	0,8	1,8	19,3	43,1	0,6	1,2
97	92,2	7,8	2,3	30,0	68,1	0,7	1,5	19,3	43,8	0,4	1,0
98	94,1	5,9	2,3	30,0	69,3	0,5	1,1	19,3	44,6	0,3	0,8
99	96,0	4,0	2,4	30,0	70,6	0,3	0,7	19,3	45,4	0,2	0,5
100	98,0	2,0	2,4	30,0	71,8	0,2	0,4	19,3	46,3	0,1	0,3
Moyenne par ZD											
	32,8	67,2	1,0		25,3		12,6		17,14		8,6
Nombre de ZD sélectionnées					23,7	ZD identiques			35,00	ZD identiques	
Taille d'échantillon attendue				nc=	600,0	ne=	300,0		600,00		300,0

Tableau 6.9: Spécification d'un nombre minimum (e_{min}) de ménages de type E à prélever dans toute zone d'échantillon

nc= 600 ne= 300 cmax= 10,0 valeur requise a= 64,7				aucune limite emin				emin= 3,0			
				Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'E' m sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
								avant l'ajustement def3		après ajustement f3 réduit par= 0,72	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
1	0,0	100,0	0,3	0,0	0,0	23,2	6,9	23,2	6,9	16,6	4,9
2	0,0	100,0	0,3	0,0	0,0	23,1	6,9	23,1	6,9	16,6	4,9
3	0,0	100,0	0,3	0,0	0,0	23,1	6,9	23,1	6,9	16,6	4,9
4	0,1	99,9	0,3	0,1	0,0	23,0	6,9	23,0	6,9	16,5	4,9
5	0,2	99,8	0,3	0,2	0,0	22,9	6,9	22,9	6,9	16,4	4,9
6	0,3	99,8	0,3	0,3	0,1	22,7	6,9	22,7	6,9	16,3	4,9
7	0,4	99,6	0,3	0,4	0,1	22,5	6,9	22,5	6,9	16,1	4,9
8	0,5	99,5	0,3	0,5	0,2	22,3	6,9	22,3	6,9	16,0	4,9
9	0,6	99,4	0,3	0,6	0,2	22,0	6,9	22,0	6,9	15,8	4,9
10	0,8	99,2	0,3	0,8	0,3	21,7	6,8	21,7	6,8	15,6	4,9
11	1,0	99,0	0,3	1,0	0,3	21,4	6,8	21,4	6,8	15,3	4,9
12	1,2	98,8	0,3	1,2	0,4	21,1	6,8	21,1	6,8	15,1	4,9
13	1,4	98,6	0,3	1,4	0,5	20,7	6,8	20,7	6,8	14,8	4,9
14	1,7	98,3	0,3	1,7	0,6	20,3	6,8	20,3	6,8	14,6	4,9
15	2,0	98,0	0,3	2,0	0,7	19,9	6,8	19,9	6,8	14,3	4,8
16	2,3	97,8	0,3	2,3	0,8	19,5	6,7	19,5	6,7	14,0	4,8
17	2,6	97,4	0,4	2,6	0,9	19,1	6,7	19,1	6,7	13,7	4,8
18	2,9	97,1	0,4	2,9	1,0	18,6	6,7	18,6	6,7	13,4	4,8
19	3,2	96,8	0,4	3,2	1,2	18,2	6,7	18,2	6,7	13,0	4,8
20	3,6	96,4	0,4	3,6	1,4	17,7	6,6	17,7	6,6	12,7	4,8
21	4,0	96,0	0,4	4,0	1,5	17,3	6,6	17,3	6,6	12,4	4,7
22	4,4	95,6	0,4	4,4	1,7	16,8	6,6	16,8	6,6	12,1	4,7
23	4,8	95,2	0,4	4,8	1,9	16,4	6,6	16,4	6,6	11,7	4,7
24	5,3	94,7	0,4	5,3	2,2	15,9	6,5	15,9	6,5	11,4	4,7
25	5,8	94,2	0,4	5,8	2,4	15,4	6,5	15,4	6,5	11,1	4,7
26	6,3	93,8	0,4	6,3	2,7	15,0	6,5	15,0	6,5	10,7	4,6
27	6,8	93,2	0,4	6,8	3,0	14,5	6,4	14,5	6,4	10,4	4,6
28	7,3	92,7	0,5	7,3	3,3	14,1	6,4	14,1	6,4	10,1	4,6
29	7,8	92,2	0,5	7,8	3,6	13,7	6,4	13,7	6,4	9,8	4,6
30	8,4	91,6	0,5	8,4	4,0	13,2	6,3	13,2	6,3	9,5	4,5
31	9,0	91,0	0,5	9,0	4,4	12,8	6,3	12,8	6,3	9,2	4,5
32	9,6	90,4	0,5	9,6	4,8	12,4	6,2	12,4	6,2	8,9	4,5
33	10,2	89,8	0,5	10,0	5,2	12,0	6,2	12,0	6,2	8,6	4,4
34	10,9	89,1	0,5	10,0	5,3	11,6	6,1	11,6	6,1	8,3	4,4
35	11,6	88,4	0,5	10,0	5,5	11,2	6,1	11,2	6,1	8,0	4,4

nc= 600 ne= 300 cmax= 10,0 valeur requise a= 64,7				aucune limite emin				emin= 3,0			
				Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'E' m sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
								avant l'ajustement def3		après ajustement f3 réduit par= 0,72	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
36	12,3	87,8	0,6	10,0	5,6	10,8	6,1	10,8	6,1	7,8	4,3
37	13,0	87,0	0,6	10,0	5,7	10,4	6,0	10,4	6,0	7,5	4,3
38	13,7	86,3	0,6	10,0	5,9	10,1	6,0	10,1	6,0	7,2	4,3
39	14,4	85,6	0,6	10,0	6,1	9,7	5,9	9,7	5,9	7,0	4,2
40	15,2	84,8	0,6	10,0	6,2	9,4	5,8	9,4	5,8	6,7	4,2
41	16,0	84,0	0,6	10,0	6,4	9,1	5,8	9,1	5,8	6,5	4,2
42	16,8	83,2	0,7	10,0	6,6	8,7	5,7	8,7	5,7	6,3	4,1
43	17,6	82,4	0,7	10,0	6,8	8,4	5,7	8,4	5,7	6,0	4,1
44	18,5	81,5	0,7	10,0	6,9	8,1	5,6	8,1	5,6	5,8	4,0
45	19,4	80,6	0,7	10,0	7,1	7,8	5,6	7,8	5,6	5,6	4,0
46	20,3	79,8	0,7	10,0	7,3	7,5	5,5	7,5	5,5	5,4	3,9
47	21,2	78,8	0,8	10,0	7,5	7,2	5,4	7,2	5,4	5,2	3,9
48	22,1	77,9	0,8	10,0	7,7	7,0	5,4	7,0	5,4	5,0	3,9
49	23,0	77,0	0,8	10,0	7,9	6,7	5,3	6,7	5,3	4,8	3,8
50	24,0	76,0	0,8	10,0	8,1	6,5	5,2	6,5	5,2	4,6	3,8
51	25,0	75,0	0,8	10,0	8,3	6,2	5,2	6,2	5,2	4,5	3,7
52	26,0	74,0	0,9	10,0	8,5	6,0	5,1	6,0	5,1	4,3	3,7
53	27,0	73,0	0,9	10,0	8,8	5,7	5,0	5,7	5,0	4,1	3,6
54	28,1	71,9	0,9	10,0	9,0	5,5	5,0	5,5	5,0	4,0	3,6
55	29,2	70,8	0,9	10,0	9,2	5,3	4,9	5,3	4,9	3,8	3,5
56	30,3	69,8	0,9	10,0	9,4	5,1	4,8	5,1	4,8	3,7	3,4
57	31,4	68,6	1,0	10,0	9,7	4,9	4,7	4,9	4,7	3,5	3,4
58	32,5	67,5	1,0	10,0	9,9	4,7	4,7	4,7	4,7	3,4	3,3
59	33,6	66,4	1,0	10,0	10,2	4,5	4,6	4,5	4,6	3,2	3,3
60	34,8	65,2	1,0	10,0	10,4	4,3	4,5	4,3	4,5	3,1	3,2
61	36,0	64,0	1,1	10,0	10,7	4,1	4,4	4,1	4,4	3,0	3,2
62	37,2	62,8	1,1	10,0	10,9	4,0	4,3	4,0	4,3	3,0	3,3
63	38,4	61,6	1,1	10,0	11,2	3,8	4,2	3,8	4,2	3,0	3,4
64	39,7	60,3	1,1	10,0	11,5	3,6	4,2	3,6	4,2	3,0	3,4
65	41,0	59,0	1,2	10,0	11,7	3,5	4,1	3,5	4,1	3,0	3,5
66	42,3	57,8	1,2	10,0	12,0	3,3	4,0	3,3	4,0	3,0	3,6
67	43,6	56,4	1,2	10,0	12,3	3,2	3,9	3,2	3,9	3,0	3,7
68	44,9	55,1	1,3	10,0	12,6	3,0	3,8	3,0	3,8	3,0	3,8
69	46,2	53,8	1,3	10,0	12,9	2,9	3,7	3,0	3,9	3,0	3,9
70	47,6	52,4	1,3	10,0	13,2	2,7	3,6	3,0	3,9	3,0	3,9
71	49,0	51,0	1,3	10,0	13,5	2,6	3,5	3,0	4,0	3,0	4,0
72	50,4	49,6	1,4	10,0	13,8	2,5	3,4	3,0	4,1	3,0	4,1

nc= 600 ne= 300 cmax= 10,0 valeur requise a= 64,7				aucune limite emin				emin= 3,0			
				Nombre de 'C' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'E' m sélectionné:		Nombre de 'E' m sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
								avant l'ajustement def3		après ajustement f3 réduit par= 0,72	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
73	51,8	48,2	1,4	10,0	14,1	2,4	3,3	3,0	4,2	3,0	4,2
74	53,3	46,7	1,4	10,0	14,4	2,2	3,2	3,0	4,3	3,0	4,3
75	54,8	45,2	1,5	10,0	14,7	2,1	3,1	3,0	4,4	3,0	4,4
76	56,3	43,8	1,5	10,0	15,0	2,0	3,0	3,0	4,5	3,0	4,5
77	57,8	42,2	1,5	10,0	15,3	1,9	2,9	3,0	4,6	3,0	4,6
78	59,3	40,7	1,6	10,0	15,7	1,8	2,8	3,0	4,7	3,0	4,7
79	60,8	39,2	1,6	10,0	16,0	1,7	2,7	3,0	4,8	3,0	4,8
80	62,4	37,6	1,6	10,0	16,3	1,6	2,6	3,0	4,9	3,0	4,9
81	64,0	36,0	1,7	10,0	16,7	1,5	2,5	3,0	5,0	3,0	5,0
82	65,6	34,4	1,7	10,0	17,0	1,4	2,4	3,0	5,1	3,0	5,1
83	67,2	32,8	1,7	10,0	17,4	1,3	2,3	3,0	5,2	3,0	5,2
84	68,9	31,1	1,8	10,0	17.7	1,2	2,1	3,0	5,3	3,0	5,3
85	70,6	29,4	1,8	10,0	18,1	1,1	2,0	3,0	5,4	3,0	5,4
86	72,3	27,8	1,8	10,0	18,4	1,0	1,9	3,0	5,5	3,0	5,5
87	74,0	26,0	1,9	10,0	18,8	1,0	1,8	3,0	5,6	3,0	5,6
88	75,7	24,3	1,9	10,0	19,2	0,9	1,7	3,0	5,7	3,0	5,7
89	77,4	22,6	2,0	10,0	19,5	0,8	1,6	3,0	5,9	3,0	5,9
90	79,2	20,8	2,0	10,0	19,9	0,7	1,4	3,0	6,0	3,0	6,0
91	81,0	19,0	2,0	10,0	20,3	0,6	1,3	3,0	6,1	3,0	6,1
92	82,8	17,2	2,1	10,0	20,7	0,6	1,2	3,0	6,2	3,0	6,2
93	84,6	15,4	2,1	10,0	21,1	0,5	1,1	3,0	6,3	3,0	6,3
94	86,5	13,5	2,1	10,0	21,5	0,4	0,9	3,0	6,4	3,0	6,4
95	88,4	11,6	2,2	10,0	21,9	0,4	0,8	3,0	6,6	3,0	6,6
96	90,3	9,8	2,2	10,0	22,3	0,3	0,7	3,0	6,7	3,0	6,7
97	92,2	7,8	2,3	10,0	22,7	0,2	0,5	3,0	6,8	3,0	6,8
98	94,1	5,9	2,3	10,0	23,1	0,2	0,4	3,0	6,9	3,0	6,9
99	96,0	4,0	2,4	10,0	23,5	0,1	0,3	3,0	7,1	3,0	7,1
100	98,0	2,0	2,4	10,0	23,9	0,1	0,1	3,0	7,2	3,0	7,2
Moyenne par ZD											
	32,8	67,2	1,0		9,3		4,6		5,7		4,6
Nombre de ZD sélectionnées					64,7		ZD identiques		ZD identiques		ZD identiques
Taille d'échantillon attendue					600,0		300,0		370,3		300,0

Tableau 6.10: Limite maximale (e_{max}) du nombre de ménages de type E à prélever dans toute zone d'échantillon

nc= 600 ne= 300 cmax= aucune limite(échantillonnage à tirage complet de la catégorie C)						Nombre de 'G' m sélectionné:					
						Calcul d'origine		Limite: échantillon<=ei		ajusté pour obtenir ne= 300	
			Nombre de 'C' m sélectionné:			Nombre sélectionné:		Nombre sélectionné:		f3 augmenté par= 1,07 Nombre sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
1	0,0	100,0	0,3	0,0	0,0	128,7	38,3	100,0	29,8	100,0	29,8
2	0,0	100,0	0,3	0,0	0,0	128,6	38,3	100,0	29,8	100,0	29,8
3	0,0	100,0	0,3	0,0	0,0	128,3	38,3	100,0	29,9	100,0	29,9
4	0,1	99,9	0,3	0,1	0,0	127,8	38,3	99,9	29,9	99,9	29,9
5	0,2	99,8	0,3	0,2	0,0	127,1	38,3	99,8	30,1	99,8	30,1
6	0,3	99,8	0,3	0,3	0,1	126,1	38,2	99,8	30,2	99,8	30,2
7	0,4	99,6	0,3	0,4	0,1	125,0	38,2	99,6	30,4	99,6	30,4
8	0,5	99,5	0,3	0,5	0,2	123,7	38,1	99,5	30,7	99,5	30,7
9	0,6	99,4	0,3	0,6	0,2	122,3	38,1	99,4	30,9	99,4	30,9
10	0,8	99,2	0,3	0,8	0,3	120,7	38,0	99,2	31,3	99,2	31,3
11	1,0	99,0	0,3	1,0	0,3	118,9	37,9	99,0	31,6	99,0	31,6
12	1,2	98,8	0,3	1,2	0,4	117,0	37,9	98,8	32,0	98,8	32,0
13	1,4	98,6	0,3	1,4	0,5	115,0	37,8	98,6	32,4	98,6	32,4
14	1,7	98,3	0,3	1,7	0,6	112,8	37,7	98,3	32,8	98,3	32,8
15	2,0	98,0	0,3	2,0	0,7	110,6	37,6	98,0	33,3	98,0	33,3
16	2,3	97,8	0,3	2,3	0,8	108,3	37,5	97,8	33,8	97,8	33,8
17	2,6	97,4	0,4	2,6	0,9	105,9	37,3	97,4	34,3	97,4	34,3
18	2,9	97,1	0,4	2,9	1,0	103,5	37,2	97,1	34,9	97,1	34,9
19	3,2	96,8	0,4	3,2	1,2	101,0	37,1	96,8	35,5	96,8	35,5
20	3,6	96,4	0,4	3,6	1,4	98,5	36,9	96,4	36,1	96,4	36,1
21	4,0	96,0	0,4	4,0	1,5	96,0	36,8	96,0	36,8	96,0	36,8
22	4,4	95,6	0,4	4,4	1,7	93,4	36,6	93,4	36,6	95,6	37,5
23	4,8	95,2	0,4	4,8	1,9	90,9	36,5	90,9	36,5	95,2	38,2
24	5,3	94,7	0,4	5,3	2,2	88,3	36,3	88,3	36,3	94,3	38,7
25	5,8	94,2	0,4	5,8	2,4	85,8	36,1	85,8	36,1	91,6	38,5
26	6,3	93,8	0,4	6,3	2,7	83,3	35,9	83,3	35,9	88,9	38,3
27	6,8	93,2	0,4	6,8	3,0	80,8	35,7	80,8	35,7	86,2	38,1
28	7,3	92,7	0,5	7,3	3,3	78,3	35,5	78,3	35,5	83,6	37,9
29	7,8	92,2	0,5	7,8	3,6	75,9	35,3	75,9	35,3	81,0	37,7
30	8,4	91,6	0,5	8,4	4,0	73,5	35,1	73,5	35,1	78,4	37,5
31	9,0	91,0	0,5	9,0	4,4	71,1	34,9	71,1	34,9	75,9	37,2
32	9,6	90,4	0,5	9,6	4,8	68,8	34,6	68,8	34,6	73,5	37,0
33	10,2	89,8	0,5	10,2	5,3	66,6	34,4	66,6	34,4	71,1	36,7
34	10,9	89,1	0,5	10,9	5,8	64,4	34,2	64,4	34,2	68,7	36,5

nc= 600 ne= 300 cmax= aucune limite(échantillonnage à tirage complet de la catégorie C)						Nombre de 'G' m sélectionné:					
						Calcul d'origine		Limite: échantillon<=ei		ajusté pour obtenir ne= 300	
				Nombre de 'C' m sélectionné:		Nombre sélectionné:		Nombre sélectionné:		f3 augmenté par= 1,07 Nombre sélectionné:	
				Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique	Si ZD dans l'échantillon	Espérance mathématique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
35	11,6	88,4	0,5	11,6	6,3	62,2	33,9	62,2	33,9	66,4	36,2
36	12,3	87,8	0,6	12,3	6,9	60,1	33,6	60,1	33,6	64,1	35,9
37	13,0	87,0	0,6	13,0	7,5	58,0	33,4	58,0	33,4	61,9	35,6
38	13,7	86,3	0,6	13,7	8,1	56,0	33,1	56,0	33,1	59,8	35,3
39	14,4	85,6	0,6	14,4	8,8	54,1	32,8	54,1	32,8	57,7	35,0
40	15,2	84,8	0,6	15,2	9,5	52,2	32,5	52,2	32,5	55,7	34,7
41	16,0	84,0	0,6	16,0	10,2	50,3	32,2	50,3	32,2	53,7	34,4
42	16,8	83,2	0,7	16,8	11,0	48,5	31,9	48,5	31,9	51,8	34,0
43	17,6	82,4	0,7	17,6	11,9	46,8	31,6	46,8	31,6	49,9	33,7
44	18,5	81,5	0,7	18,5	12,8	45,1	31,2	45,1	31,2	48,1	33,3
45	19,4	80,6	0,7	19,4	13,8	43,4	30,9	43,4	30,9	46,3	33,0
46	20,3	79,8	0,7	20,3	14,8	41,8	30,6	41,8	30,6	44,6	32,6
47	21,2	78,8	0,8	21,2	15,9	40,3	30,2	40,3	30,2	43,0	32,3
48	22,1	77,9	0,8	22,1	17,0	38,8	29,9	38,8	29,9	41,4	31,9
49	23,0	77,0	0,8	23,0	18,2	37,3	29,5	37,3	29,5	39,8	31,5
50	24,0	76,0	0,8	24,0	19,5	35,9	29,1	35,9	29,1	38,3	31,1
51	25,0	75,0	0,8	25,0	20,8	34,5	28,7	34,5	28,7	36,9	30,7
52	26,0	74,0	0,9	26,0	22,2	33,2	28,4	33,2	28,4	35,4	30,3
53	27,0	73,0	0,9	27,0	23,7	31,9	28,0	31,9	28,0	34,1	29,8
54	28,1	71,9	0,9	28,1	25,2	30,7	27,6	30,7	27,6	32,7	29,4
55	29,2	70,8	0,9	29,2	26,9	29,5	27,2	29,5	27,2	31,4	29,0
56	30,3	69,8	0,9	30,3	28,6	28,3	26,7	28,3	26,7	30,2	28,5
57	31,4	68,6	1,0	31,4	30,4	27,2	26,3	27,2	26,3	29,0	28,1
58	32,5	67,5	1,0	32,5	32,3	26,1	25,9	26,1	25,9	27,8	27,6
59	33,6	66,4	1,0	33,6	34,2	25,0	25,4	25,0	25,4	26,7	27,1
60	34,8	65,2	1,0	34,8	36,3	24,0	25,0	24,0	25,0	25,6	26,7
61	36,0	64,0	1,1	36,0	38,4	23,0	24,5	23,0	24,5	24.5	26,2
62	37,2	62,8	1,1	37,2	40,7	22,0	24,1	22,0	24,1	23,5	25,7
63	38,4	61,6	1,1	38,4	43,0	21,1	23,6	21,1	23,6	22,5	25,2
64	39,7	60,3	1,1	39,7	45,5	20,2	23,1	20,2	23,1	21,5	24,7
65	41,0	59,0	1,2	41,0	48,1	19,3	22,6	19,3	22,6	20,6	24,2
66	42,3	57,8	1,2	42,3	50,8	18,4	22,1	18,4	22,1	19,7	23,6
67	43,6	56,4	1,2	43,6	53,6	17,6	21,6	17,6	21,6	18,8	23,1
68	44,9	55,1	1,3	44,9	56,5	16,8	21,1	16,8	21,1	17,9	22,5
69	46,2	53,8	1,3	46,2	59,5	16,0	20,6	16,0	20,6	17,1	22,0
70	47,6	52,4	1,3	47,6	62,7	15,3	20,1	15,3	20,1	16,3	21,4

nc= 600 ne= 300 cmax= aucune limite(échantillonnage à tirage complet de la catégorie C)						Nombre de 'G' m sélectionné:					
						Calcul d'origine		Limite: échantillon<=ei		ajusté pour obtenir ne= 300	
				Nombre de 'C' m sélectionné:		Nombre sélectionné:		Nombre sélectionné:		f3 augmenté par= 1,07 Nombre sélectionné:	
				Si ZD dans l'échan-tillon	Espérance mathé-matique	Si ZD dans l'échan-tillon	Espérance mathé-matique	Si ZD dans l'échan-tillon	Espérance mathé-matique	Si ZD dans l'échan-tillon	Espérance mathé-matique
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
i	ci	ei	Mi	ci'	ci' * Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi	ei*f3i	ei*f3i* Mi
71	49,0	51,0	1,3	49,0	65,9	14,5	19,5	14,5	19,5	15,5	20,9
72	50,4	49,6	1,4	50,4	69,4	13,8	19,0	13,8	19,0	14,7	20,3
73	51,8	48,2	1,4	51,8	72,9	13,1	18,5	13,1	18,5	14,0	19,7
74	53,3	46,7	1,4	53,3	76,6	12,5	17,9	12,5	17,9	13,3	19,1
75	54,8	45,2	1,5	54,8	80,4	11,8	17,3	11,8	17,3	12,6	18,5
76	56,3	43,8	1,5	56,3	84,4	11,2	16,8	11,2	16,8	11,9	17,9
77	57,8	42,2	1,5	57,8	88,5	10,6	16,2	10,6	16,2	11,3	17,3
78	59,3	40,7	1,6	59,3	92,8	10,0	15,6	10,0	15,6	10,6	16,7
79	60,8	39,2	1,6	60,8	97,3	9,4	15,0	9,4	15,0	10,0	16,0
80	62,4	37,6	1,6	62,4	101,9	8,8	14,4	8,8	14,4	9,4	15,4
81	64,0	36,0	1,7	64,0	106,7	8,3	13,8	8,3	13,8	8,8	14,7
82	65,6	34,4	1,7	65,6	111,6	7,7	13,2	7,7	13,2	8,3	14,1
83	67,2	32,8	1,7	67,2	116,7	7,2	12,6	7,2	12,6	7,7	13,4
84	68,9	31,1	1,8	68,9	122,0	6,7	11,9	6,7	11,9	7,2	12,7
85	70,6	29,4	1,8	70,6	127,5	6,2	11,3	6,2	11,3	6,7	12,0
86	72,3	27,8	1,8	72,3	133,2	5,8	10,6	5,8	10,6	6,2	11,4
87	74,0	26,0	1,9	74,0	139,0	5,3	10,0	5,3	10,0	5,7	10,7
88	75,7	24,3	1,9	75,7	145,1	4,9	9,3	4,9	9,3	5,2	9,9
89	77,4	22,6	2,0	77,4	151,3	4,4	8,6	4,4	8,6	4,7	9,2
90	79,2	20,8	2,0	79,2	157,8	4,0	8,0	4,0	8,0	4,3	8,5
91	81,0	19,0	2,0	81,0	164,4	3,6	7,3	3,6	7,3	3,8	7,8
92	82,8	17,2	2,1	82,8	171,3	3,2	6,6	3,2	6,6	3,4	7,0
93	84,6	15,4	2,1	84,6	178,4	2,8	5,9	2,8	5,9	3,0	6,3
94	86,5	13,5	2,1	86,5	185,7	2,4	5,2	2,4	5,2	2,6	5,5
95	88,4	11,6	2,2	88,4	193,3	2,0	4,5	2,0	4,5	2,2	4,8
96	90,3	9,8	2,2	90,3	201,1	1,7	3,7	1,7	3,7	1,8	4,0
97	92,2	7,8	2,3	92,2	209,1	1,3	3,0	1,3	3,0	1,4	3,2
98	94,1	5,9	2,3	94,1	217,4	1,0	2,3	1,0	2,3	1,0	2,4
99	96,0	4,0	2,4	96,0	225,9	0,6	1,5	0,6	1,5	0,7	1,6
100	98,0	2,0	2,4	98,0	234,6	0,3	0,8	0,3	0,8	0,3	0,8
Moyenne par ZD			32,8	32,8	51,5						25,7
Nombre de ZD sélectionnée					11,7						11,7
Taille d'échantillon attendue					600,0		300,0		286,3		300,0

Tableau 6.11: Exemple d'échantillonnage systématique circulaire avec une probabilité proportionnelle à la taille

Paramètres d'échantillonnage

nc= 600
ne= 300

Base d'échantillonnage ZD			Mesure de taille (MoS)	Cumulé MoS	Point de sélection	Ordre de sélection
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	ci	ei	Mi	Ti	Si	
1	0,0	100,0	0,30	0,30		
2	0,0	100,0	0,30	0,60		
3	0,0	100,0	0,30	0,89		
4	0,1	99,9	0,30	1,19		
5	0,2	99,8	0,30	1,50	1,27	9
6	0,3	99,8	0,30	1,80		
7	0,4	99,6	0,31	2,10		
8	0,5	99,5	0,31	2,41		
9	0,6	99,4	0,31	2,72		
10	0,8	99,2	0,32	3,04		
11	1,0	99,0	0,32	3,36		
12	1,2	98,8	0,32	3,68		
13	1,4	98,6	0,33	4,01		
14	1,7	98,3	0,33	4,34		
15	2,0	98,0	0,34	4,68		
16	2,3	97,8	0,35	5,03		
17	2,6	97,4	0,35	5,38		
18	2,9	97,1	0,36	5,74		
19	3,2	96,8	0,37	6,11	5,82	10
20	3,6	96,4	0,37	6,48		
21	4,0	96,0	0,38	6,87		
22	4,4	95,6	0,39	7,26		
23	4,8	95,2	0,40	7,66		
24	5,3	94,7	0,41	8,07		
25	5,8	94,2	0,42	8,49		
26	6,3	93,8	0,43	8,92		
27	6,8	93,2	0,44	9,37		
28	7,3	92,7	0,45	9,82		
29	7,8	92,2	0,47	10,29		
30	8,4	91,6	0,48	10,76	10,36	11
31	9,0	91,0	0,49	11,25		
32	9,6	90,4	0,50	11,76		
33	10,2	89,8	0,52	12,27		
34	10,9	89,1	0,53	12,80		
35	11,6	88,4	0,55	13,35		
36	12,3	87,8	0,56	13,91		
37	13,0	87,0	0,57	14,48		
38	13,7	86,3	0,59	15,07	14,91	12

Paramètres d'échantillonnage

nc= 600
ne= 300

Base d'échantillonnage ZD			Mesure de taille (MoS)	Cumulé MoS	Point de sélection	Ordre de sélection
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	ci	ei	Mi	Ti	Si	
39	14,4	85,6	0,61	15,68		
40	15,2	84,8	0,62	16,30		
41	16,0	84,0	0,64	16,94		
42	16,8	83,2	0,66	17,60		
43	17,6	82,4	0,68	18,28		
44	18,5	81,5	0,69	18,97		
45	19,4	80,6	0,71	19,68	19,46	13
46	20,3	79,8	0,73	20,41		
47	21,2	78,8	0,75	21,16		
48	22,1	77,9	0,77	21,93		
49	23,0	77,0	0,79	22,72		
50	24,0	76,0	0,81	23,53		
51	25,0	75,0	0,83	24,37	24,00	14
52	26,0	74,0	0,85	25,22		
53	27,0	73,0	0,88	26,10		
54	28,1	71,9	0,90	27,00		
55	29,2	70,8	0,92	27,92		
56	30,3	69,8	0,94	28,86	28,55	15
57	31,4	68,6	0,97	29,83		
58	32,5	67,5	0,99	30,82		
59	33,6	66,4	1,02	31,84		
60	34,8	65,2	1,04	32,88		
61	36,0	64,0	1,07	33,95	33,09	16
62	37,2	62,8	1,09	35,04		
63	38,4	61,6	1,12	36,16		
64	39,7	60,3	1,15	37,31		
65	41,0	59,0	1,17	38,48	37,64	17
66	42,3	57,8	1,20	39,68		
67	43,6	56,4	1,23	40,91		
68	44,9	55,1	1,26	42,17		
69	46,2	53,8	1,29	43,46	42,18	18
70	47,6	52,4	1,32	44,77		
71	49,0	51,0	1,35	46,12		
72	50,4	49,6	1,38	47,50	46,73	19
73	51,8	48,2	1,41	48,90		
74	53,3	46,7	1,44	50,34		
75	54,8	45,2	1,47	51,81	51,27	20
76	56,3	43,8	1,50	53,31		
77	57,8	42,2	1,53	54,84		
78	59,3	40,7	1,57	56,41	55,82	21

Paramètres d'échantillonnage

nc= 600
ne= 300

Base d'échantillonnage ZD			Mesure de taille (MoS)	Cumulé MoS	Point de sélection	Ordre de sélection
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	ci	ei	Mi	Ti	Si	
79	60,8	39,2	1,60	58,01		
80	62,4	37,6	1,63	59,64		
81	64,0	36,0	1,67	61,31	60,36	22
82	65,6	34,4	1,70	63,01		
83	67,2	32,8	1,74	64,74		
84	68,9	31,1	1,77	66,51	64,91	1
85	70,6	29,4	1,81	68,32		
86	72,3	27,8	1,84	70,16	69,46	2
87	74,0	26,0	1,88	72,04		
88	75,7	24,3	1,92	73,96		
89	77,4	22,6	1,95	75,91	74,00	3
90	79,2	20,8	1,99	77,91		
91	81,0	19,0	2,03	79,94	78,55	4
92	82,8	17,2	2,07	82,00		
93	84,6	15,4	2,11	84,11	83,09	5
94	86,5	13,5	2,15	86,26		
95	88,4	11,6	2,19	88,45	87,64	6
96	90,3	9,8	2,23	90,68		
97	92,2	7,8	2,27	92,94	92,18	7
98	94,1	5,9	2,31	95,25		
99	96,0	4,0	2,35	97,61	96,73	8
100	98,0	2,0	2,39	100,00		
Moyenne par ZD						
	32,8	67165,0				
Nombre de ZD sélectionnées						
Taille d'échantillon attendue						

Chapitre 7

Estimation fondée sur les données d'échantillon

Le présent chapitre aborde des questions précises liées à l'estimation fondée sur les données d'enquête sur échantillon. Comme au chapitre 3, cette estimation est réalisée dans une perspective globale pour pouvoir être utile dans toute enquête de population (notamment les EMO), bien que le contexte supposé soit en fait celui des enquêtes sur le travail des enfants.

Nous examinerons d'abord les questions et les procédures statistiques liées à la pondération des données d'échantillon et la production des estimations tirées des enquêtes (Sections 7.4-7.7.3). Ensuite, nous abordons les procédures pratiques de traitement et d'analyse de l'information appropriée sur les erreurs d'échantillonnage pour une application à grande échelle dans de vastes enquêtes complexes (Sections 7.4-7.7). Enfin, des implications sont tirées par rapport à une question qui revêt une grande valeur pratique – la détermination du nombre approprié de grappes (ou le prélévement de l'échantillon par grappe, pour une taille d'échantillon donnée).

7.1. Pondération des données d'échantillon

Dans la plupart des cas (même si ce n'est pas nécessairement dans tous), les données d'échantillon doivent être pondérées afin de produire des estimations pour la population d'intérêt. Quand les données d'échantillon *doivent être pondérées*, il est fortement souhaitable – pour des besoins de commodité pratique – *d'associer chaque cas ou dossier à son poids en tant que variable du fichier de microdonnées*. La plupart des estimations requises, tels que les proportions, les moyennes, les ratios, les taux, etc., peuvent ainsi être produits de manière très directe sans se référer davantage à la structure de l'échantillon. L'estimation de la variance gagne aussi en simplicité: la plupart des méthodes de traitement des erreurs d'échantillonnage nécessitent simplement des ensembles pondérés au niveau des unités de sondage du premier degré, l'identification des UEP et des strates dans lesquelles elles se trouvent comme défini afin de traiter les erreurs d'échantillonnage.

Toutefois, il est nécessaire de partir d'une question très élémentaire qui se pose dans la production des estimations tirées d'une enquête: *à savoir si les données de l'échantillon doivent être pondérées ou non.* La pondération des données d'échantillon a pour objectif d'améliorer la représentativité de l'échantillon en termes de taille, de distribution et de caractéristiques de la population d'étude. Lorsque les unités d'échantillon ont été sélectionnées avec des probabilités différentes, par exemple, il est nécessaire de pondérer les résultats de manière inversement proportionnelle, aux probabilités de sélection des unités, afin de refléter la situation réelle au sein de la population. Une enquête suffissament bien conçue et mise en œuvre, engendrant des

taux de réponses élévés requiert l'application du «systéme de pondération» susmentionné.

Cependant, dans la pratique, la situation est souvent plus complexe en raison des défaillances dans la sélection et la mise en oeuvre de l'échantillon (ce qui biaise les résultats) ainsi que de la necessité (et de la possibilité) de réduire les variances par des procédés d'estimations améliorés. Le besoin en procédures d'estimation et en pondération plus complexes tend à être plus élevé dans les enquêtes qui présentent un fort taux de non réponses et d'erreurs de couverture, des inconsistances dans les définitions des unités utilisées lors des diverses étapes au cours de l'enquête, des divergences avec l'échantillonnage représentatif (probabiliste), des petites tailles d'échantillon; et lorsque les estimations de plusieurs sous-populations distinctes doivent être produites. Le besoin (et l'opportunité) est aussi plus élevé en présence d'informations externes plus appropriées et plus fiables par rapport à l'objectif visé. Ces facteurs doivent être pris en considération lorsqu'une enquête sur le travail des enfants est menée.

7.1.1. Échantillons autopondérés

Un échantillon autopondéré signifie que chaque unité élémentaire de la population a la même probabilité non nulle d'entrer dans l'échantillon. Les unités des étapes plus élevées peuvent bien sûr être choisies avec des probabilités différentes, mais de telle sorte que ces différences de probablités de prélévement aux étapes diverses se compensent. Avec des échantillons autopondérés, les estimations d'échantillon peuvent être préparées à partir de données non pondérées. Les résultats sont alors multipliés, si nécessaire, par un facteur général constant. Certains facteurs expliquent la valeur des plans d'autopondération.

- La pondération accroît la complexité de l'analyse de l'enquête.
- Des poids aléatoires, qui ne sont pas liés aux variances de la population, augmentent la variance des résultats.
- La pondération peut réduire la flexibilité et l'aisance avec laquelle le même échantillon peut être utilisé à des fins diverses et à des enquêtes différentes. Pour satisfaire aux conditions des divers sujets et enquêtes, le compromis d'une allocation à buts multiples peut se rapprocher de l'autopondération.
- Les échantillons autopondérés sont plus facilement assimilés et acceptés par l'utilisateur non statisticien et le grand public.
- Il faut noter que les divergences modérées par rapport à l'autopondération ont un léger effet sur les variances. En d'autres termes, le sur-échantillonnage de l'allocation optimale ou la pondération à d'autres fins, au stade d'analyse, n'est valable que s'il implique une divergence relativement importante par rapport à l'autopondération.

Toutefois, l'on peut avoir de bonnes raisons de faire diverger les plans d'autopondération:

- Si les considérations ci-dessus en faveur de l'autopondération s'appliquent en particulier aux enquêtes à grande échelle sur les ménages de la population

générale, les enquêtes visant à représenter séparément certaines sous-populations (tels que les enfants travailleurs dans les divers secteurs d'activités) ont souvent besoin de répartitions disproportionnées comme le sur-échantillonnage de petites sous-populations d'intérêt.

- La divergence avec l'autopondération est souvent requise dans les enquêtes sur les enfants travailleurs pour atteindre la cible de la taille d'échantillon et s'assurer un nombre suffisant d'enfants travailleurs dans chaque zone d'échantillon. Cela est dû au fait que la population cible est relativement rare et inégalement répartie.
- Même dans des enquêtes générales sur les ménages, il est parfois nécessaire d'utiliser des taux d'échantillonnage differents dans les domaines différents, par exemple pour représenter de façon adéquate les domaines plus petits, plus importants ou plus variables.
- Des contraintes pratiques comme le besoin d'avoir une taille d'échantillon fixe dans chaque zone de dénombrement peut aboutir à diverses probabilités de sélection.
- Même lorsque, dans la conception, les éléments sont sélectionnés avec des probabilités égales, tant dans la sélection que dans la mise en œuvre, l'échantillon résultant peut toujours se révéler être un échantillon non autopondéré, en raison de la défectuosité dans la base, des erreurs de sélection, des non-réponses, etc.
- De nombreuses enquêtes sont inadaptées pour produire de bonnes estimations d'ensembles de population (par opposition aux moyennes et aux proportions), sans pondération avec les normes externes. Par ailleurs, un système de pondération cohérent permet d'assurer la cohérence des estimations de sources différentes.

7.1.2. Quand pondérer ?

Les décisions pour pondérer les données d'échantillon exigent un équilibre entre divers coûts et avantages de la pondération. Les coûts englobent la complexité accrue, l'inconvenance, le travail de programmation et d'analyse, le coût, les possibilités d'erreurs ou la mauvaise utilisation des données, une variance avec une pondération aléatoire par commodité et même un accroissement des biais si des normes inadaptées sont utilisées pour modifier (re-pondérer) les résultats de l'enquête. Parmi les avantages, on citerait la réduction des biais et éventuellement la réduction de la variance qui peut être réalisée au moyen de procédures d'estimation plus élaborées. En pratique, *la pondération devrait être envisagée quand les divergences par rapport à l'autopondération sont considérables en raison de l'effet combiné de probabilités d'échantillonnage différentes de la défectuosité de la base de sondage, des non réponses ou autres (c'est-à-dire à l'extérieur de l'intervalle +/-20%)*. Il n'est généralement pas profitable d'appliquer des poids différents quand l'étendue de leurs variations est plus réduite.

7.1.3. Effet des poids sur la variance et le biais

Il existe de nombreuses situations dans lesquelles différentes parties de la population sont échantillonnées à différents taux, déterminés par certaines conditions (par exemple, la production des estimations avec une précision minimale indiquée pour les petits domaines). Toutefois, les variations des probabilités de sélection et les poids d'échantillon associés sont indépendents des variances, des coûts et d'autres caractéristiques de la population. Dans ce sens, les poids requis pour compenser un tel différentiel de taux d'échantillonnage peuvent être considérés comme arbitraires ou choisis au hasard. Leur effet consiste généralement à accroître la variance dans un échantillon d'autopondération correspondant. Une approximation proche du facteur par lequel la variance augmente est la suivante:

$$D_W^2 = \frac{n * \Sigma_s w_i^2}{(\Sigma_s w_i)^2} = \left(1 + cv^2(w_i)\right)$$

où $cv(w_i)$ est le coefficient de variation des poids individuels w_i et la somme concerne les unités n de l'échantillon.

La signification de la formule ci-dessus consiste en ce qu'elle donne l'ampleur de l'effet par lequel *toutes les variances pour les estimations d'enquête différentes (diverses variables dans différentes sous-classes, comparaisons entre les sous-classes) sont augmentées de manière plus ou moins uniforme* suite à une pondération aléatoire. Cela constitue l'utilité pratique de l'isolement de cet effet.

Le *biais* résultant du fait d'ignorer les poids dépend de la différence à la fois dans les valeurs moyennes et les tailles des groupes ayant des poids différents, et il n'est pas le même pour des types différents de statistiques. Par conséquent, l'ampleur relative du biais par rapport à l'effet de la pondération sur la variance peut varier selon le type de statistique considéré. Malheureusement, pour cette raison, nous ne sommes pas capables de donner des conseils précis sur l'équilibre indiqué entre l'accroissement éventuel de la variance et la réduction du biais suite à l'introduction des poids d'échantillon.

Eviter les poids extrêmes

Toutefois, une chose est claire. Le souci fondamental dans la détermination du choix des poids est de maximiser leur contribution dans la réduction de l'erreur totale due à la variance et au biais dans les estimations résultantes. En pratique, il devient souhaitable d'éviter l'introduction de poids extrêmes, surtout les poids très élevés. L'utilisation de (grands) poids extrêmement variables, même lorsqu'ils affectent seulement une petite partie des cas d'échantillon, peut aboutir à une augmentation substantielle de la variance, alors même que leur contribution à la réduction du biais est peut-être faible.

Il est par conséquent courant d'ajuster les poids extrêmes à un intervalle indiqué de manière à limiter la croissance associée de la variance. Bien que des approches assez subtiles soient possibles, de nombreuses organisations estiment approprié de se fier aux règles simples pour ajuster les valeurs extrêmes de poids, tout au moins pour la production courantes d'informations statistiques. Une recommandation pratique est

que, sauf pour les poids du plan, les poids extrêmes devraient être réduits pour que *le ratio du plus grand sur le cas du plus petit introduit ne dépasse pas 5 ou à peu près*.

7.2. Calcul des poids d'échantillon: une approche systématique

Les questions liées à la pondération de l'échantillon sont généralement complexes et la «meilleure» solution peut être spécifique à la situation, selon la nature des données disponibles, les sources d'erreurs qui doivent être contrôlées, la connaissance des circonstances spécifiques et les limites de l'enquête ainsi que les pratiques existantes et les préférences dans l'organisation de l'enquête. Néanmoins, il y a de nombreux avantages à suivre certaines normes de base et une approche systématique.

7.2.1. Sources d'information sur la pondération

Dans l'application des poids, une meilleure utilisation doit être faite des informations disponibles, internes à l'échantillon ainsi que des sources externes. On donne le rôle principal aux informations internes à l'enquête; les informations externes sont présentées selon la mesure jugée utile pour une amélioration de la représentativité de l'échantillon. Divers types de sources d'information peuvent être utilisés de façon systématique pour appliquer des poids dans une procédure étape par étape:

- Le plan d'échantillonnage, c'est-à-dire les probabilités du plan de sélection de chaque unité finale (comme le ménage).
- La base d'échantillonnage qui peut fournir des informations complémentaires sur les secteurs d'échantillonnage et sur toutes les unités de réponses et de non réponses.
- La mise en œuvre de l'échantillon, c'est-à-dire le taux de réponses et d'informations sur les non enquêtés.
- D'autres enquête(s) significativement plus vaste(s) ayant une meilleure couverture, un taux de réponse élevé et des informations plus fiables sur certaines caractéristiques des ménages et/ou des personnes. A titre d'exemple, l'utilisation d'une enquête récente à grande échelle avec une bonne couverture et un taux de réponse élevé (notamment les EMO de nombreux pays) pour améliorer la représentativité d'une enquête plus complexe et difficile avec une taille d'échantillon plus petite (comme une enquête sur les enfants travailleurs).
- Les registres actuels ou les projections de la population du recensement, fournissant des informations sur la taille, les caractéristiques et la répartition de la population.

Lorsqu'une même information ou une information semblable est disponible dans plus d'une source, on devrait accorder la priorité à la source interne à l'enquête. *La pondération pour compenser les différences de probabilités de sélection et l'incidence connue de la non-réponse, par exemple, devrait toujours être appliquée avant l'introduction des corrections sur la base des données externes*. Dans

l'utilisation des informations externes, il faut s'assurer qu'elles sont significativement plus fiables que celles qui sont disponibles à l'intérieur de l'enquête, que les éléments d'information utilisés sont comparativement définis et mesurés dans les deux sources et que la couverture et le champ d'application sont les mêmes. Par exemple, si l'enquête est limitée à la population résidant dans les ménages privés, l'information externe sur la population utilisée pour la pondération de l'échantillon devrait être limitée en conséquence.

En utilisant l'information externe comme la norme d'ajustement des données de l'enquête, les facteurs susmentionnés indiquent que l'information externe fondée sur les enquêtes sur échantillon devrait avoir la primauté sur l'information semblable tirée d'autres types de sources comme les documents et les registres administratifs. Il en est ainsi parce que les données provenant de sources de même type sont généralement plus compatibles par rapport aux données issues de différents types de sources.

Les remarques ci-dessus s'appliquent aux statistiques comme les taux, les ratios et les répartitions estimés à partir de l'échantillon. De façon générale, les enquêtes sur échantillon sont nettement mieux dans l'estimation de telles statistiques qu'à celle de l'ensemble de la population. Pour cette dernière, les sources alternatives comme le recensement, les projections de population et éventuellement les registres et d'autres documents administratifs sont souvent plus fiables.Voir Section 7.3.

Procédure étape par étape

Pour respecter les normes courantes et par souci de clarté et de commodité, il est souhaitable qu'une procédure systématique soit adoptée, séparant les différents aspects de la pondération. En règle générale, chaque étape devrait être appliquée séparément, pour que sa contribution aux poids finals puisse être identifiée. Les sous-sections suivantes décrivent les étapes principales.

7.2.2. Poids du plan

Les poids du plan sont introduits pour compenser les différences de probabilités de sélection dans l'échantillon. *Chaque unité finale de l'échantillon est pondérée par la proportion inverse à la probabilité avec laquelle elle a été sélectionnée.*

Les valeurs relatives importent davantage que les valeurs absolues des probabilités de sélection. Avec un plan d'échantillonnage à plusieurs degrés, on se réfère aux probabilités de sélection générales des ménages. Dans de nombreuses enquêtes auprès des ménages, les unités finales sont sélectionnées avec des probabilités uniformes, de sorte que les poids du plan sont identiques (ex. $=1.0$). Dans d'autres cas, les échantillons ont peut-être été conçus pour être autopondérées au sein de l'échantillonnage principal ou des domaines de rapport, avec des taux de sélection variables à travers les domaines; dans ces cas, toutes les unités qui sont à l'intérieur d'un domaine donné reçoivent le même poids du plan, et les poids varient à travers les domaines par proportion inverse aux probabilités de sélection du domaine.

Si p_i est la probabilité générale d'échantillonnage du ménage i (et de toutes les personnes de ce ménage) et n, le nombre de ménages dénombrés avec succès dans l'échantillon, les poids du plan seront:

$$w_i = \left(\frac{n}{\sum(1/p_i)} \right) . \frac{1}{p_i}$$

où le facteur entre parenthèse est simplement une constante choisie pour s'assurer que la valeur moyenne du poids par ménage (dénombré avec succès dans l'enquête) est 1,0. Une telle échelle est souvent appropriée dans la pratique.

7.2.3. Pondération de l'erreur de couverture

Dans certaines circonstances, il est utile (et nécessaire) d'intégrer une correction aux poids du plan pour *l'exclusion ou la sous couverture brute* connus de certaines parties de la population d'étude. Cela peut se produire à cause des insuffisances dans la base de sondage ou pour d'autres raisons liées aux procédures de sélection et de mise en œuvre de l'échantillonnage. Un moyen d'appliquer une telle correction consisterait à réduire les probabilités du plan de sélection (en d'autres termes, pour accroître les poids correspondants) proportionnellement au taux de couverture des domaines affectés ou pour intégrer la compensation dans d'autres domaines couverts semblables à ceux exclus. *Cela peut se révéler important dans les enquêtes sur le travail des enfants dans lesquelles les bases de sondages disponibles pour la sélection de l'échantillon sont peut-être incomplètes ou dans lesquelles il serait impossible de dénombrer certaines parties de la population.*

Il est préférable d'estimer (même vaguement) le degré d'incomplétude et de l'intégrer comme un ajustement de l'estimation plutôt que de l'ignorer entiérement.

Voici un exemple simple illustrant ce point.

Supposez que deux des différents domaines de l'enquête sont constitués de très petites unités de zones: (1) la plus petite et (2) la plus grande proche. Suppossons que le domaine (1) représente 2% de la population totale mais qu'à cause du coût et des raisons pratiques, ce domaine ne peut pas être inclus dans l'échantillon. Supposons que le domaine (2) représente 3% de la population totale. Ce domaine est celui qui ressemble le plus au domaine (1). De ce fait, il a été décidé d'augmenter son poids pour compenser la non inclusion de (1). Pour représenter le total de 5% de ces deux domaines, les poids du plan dans (2) doivent tous être multipliés par le facteur (5/3).

7.2.4. Les poids de non-réponse

Ils sont introduits pour réduire l'effet des différences de taux de réponse réalisés dans différentes parties de l'échantillon. Ces poids peuvent seulement être estimés par rapport aux caractéristiques connues des unités de réponse et de non réponse. La pondération des non-réponses est particulièrement importante lorsque les taux de non-réponse sont élevés et variables de part et d'autre de l'échantillon.

La pondération des non-réponses implique la division de l'échantillon en des «classes de pondération» appropriées et, dans chaque classe de pondération – en essayant de «compenser» les cas de non réponse dans cette classe de pondération – la pondération

des unités de réponse correspondantes dans la *proportion inverse au taux de réponse de la classe*.

Bien évidemment, la pondération des non-réponses ne peut pas prendre en compte l'effet des niveaux absolus de non réponse – au mieux, elle corrige seulement les non-réponses différentielles entre les classes. L'efficacité de la procédure dépend de la mesure dans laquelle les unités de non réponse de chaque classe sont semblables aux unités de réponse de cette classe, en ce qui concerne les variables importantes. Les différences des caractéristiques d'unités et de taux de réponse devraient être maximisés à travers des classes de pondération choisies. Il est manifeste que les classes de pondération ne peuvent être définies que si les caractéristiques sont disponibles et pour les unités de réponse et pour les unités de non-réponse.

Étant donné cette condition, il est toujours nécessaire de choisir le nombre et la taille appropriés des classes à utiliser à cette fin. L'utilisation de nombreuses classes de pondération comporte l'avantage éventuel de réduire le biais de non-réponse en créant des catégories de pondération relativement petites et homogènes au sein desquelles les caractéristiques des enquêtés et des non enquêtés peuvent être considérées comme semblables. Parallèlement, l'utilisation de nombreuses petites classes de pondération peut aboutir à l'application de poids élevés et variables qui peuvent accroître considérablement la variance des estimations de l'échantillon. Un compromis est donc requis. Le choix dépendra de la variabilité des taux de réponse à travers différentes parties de l'échantillon et de la façon dont ces variations sont liées aux caractéristiques des unités.

Les classes de pondération ayant une taille moyenne d'au *moins 100* cas d'échantillon peuvent souvent être un choix raisonnable. Dans un échantillon de 5 000 ménages par exemple, cela revient à créer 50 ou moins encore de classes de pondération.

Les classes de pondération ne peuvent être définies que sur les caractéristiques disponibles et pour les unités de réponse et pour les unités de non-réponse. Dans un échantillon de ménages à plusieurs degrés, les caractéristiques impliquées peuvent être de deux types: celles qui appartiennent aux zones d'échantillon et celles qui concernent chaque ménage individuel, qu'il s'agissent de ménages enquêtés ou non enquêtés.

Les caractéristiques du niveau de zone renvoient aux caractéristiques qui ont trait aux zones ou à d'autres ensembles, comme le lieu géographique (circonscription administrative), le type de place de résidence (classification urbaine/rurale) et diverses caractéristiques socioéconomiques de ces zones. De telles informations sont toujours disponibles dans le plan même de l'échantillon (précisément, la situation géographique et d'autres informations utilisées pour la stratification des zones d'échantillonnage). D'autres informations peuvent aussi provenir de la base de sondage ou des sources externes comme les statistiques du recensement de la localité ou des sources administratives.

Les caractéristiques du ménage renvoient aux caractéristiques liées à chaque ménage tels que le mode d'occupation du logement, la taille et le type de ménage, le statut socioéconomique et d'autres caractéristiques du chef de ménage et/ou personne de référence, le nombre de personnes qui travaillent et d'autres caractéristiques qui

peuvent être liées au niveau de vie du ménage, au revenu et à l'activité économique. Nous pouvons aussi mentionner les caractéristiques similaires au niveau personnel. Diverses sources de ces informations sont disponibles. Dans les cas où l'échantillon est tiré des listes qui renferment des informations pertinentes pour la classification des unités sélectionnées, les informations de la base de sondage peuvent être utilisées pour fournir des classifications courantes des unités de réponse et de non-réponse. Dans certains cas, une source d'information supplémentaire peut être le couplage des unités de l'échantillon avec des sources externes comme le recensement de la population ou les documents administratifs par lesquels il est nécessairement possible d'accéder aux données du micro-niveau. Lorsque de telles informations externes sont utilisées à cette fin, la même source devrait être utilisée pour la classification des ménages enquêtés et non enquêtés, même si pour les derniers le même type d'information est aussi disponible dans l'enquête elle-même. Cela est nécessaire pour maintenir la cohérence de la classification. Dans les cas où l'échantillon du moment est basé sur une enquête précédente, cette dernière peut fournir des informations sur les personnes non interrogées de l'enquête en cours. Enfin, on peut aussi mentionner que, surtout dans les enquêtes complexes prédisposées aux taux élevés de non-réponse, il peut être utile de fournir un effort particulier pour collecter au moins les éléments d'information de base sur chaque unité sélectionnée dans l'échantillon, sans tenir compte du fait que l'unité est dénombrée ou non avec succès dans l'enquête principale.

En principe, le calcul du taux de réponse est direct: c'est le ratio entre le nombre d'unités dénombrées avec succès et le nombre sélectionné au départ. Par exemple, dans un échantillon de ménages, le taux de réponse R est calculé comme le ratio entre le nombre de ménages interrogés (m) et le nombre sélectionné (n) dans la classe de pondération:

$$R_j = m_j / n_j$$

Le poids de non-réponse requis est

$$w_j = \overline{R} / R_j$$

où l'inverse des valeurs R_j ont été multipliées par leur moyenne de manière à élever les poids à 1,0 en moyenne, pour plus de commodité.

Dans la pratique, les complications peuvent survenir pour diverses raisons. Tout d'abord, le type d'unité d'échantillonnage peut ne pas être identique au type d'unités dénombrées au final dans l'enquête. Par exemple, nous pouvons sélectionner un échantillon d'habitations pour les ménages dénombrés ou un échantillon d'habitations ou de ménages pour le dénombrement des personnes. Dans de telles situations le nombre exact d'unités de dénombrement sélectionnées (dénominateur du taux de réponse) peut être inconnu.

Deuxièmement, dans le calcul réel, il faut prendre soin de bien définir le dénominateur dans l'expression ci-dessus. Il devrait intégrer uniquement les unités valables, notamment les adresses contenues dans l'échantillon, c'est-à-dire exclure les adresses vides, inexistantes, inaccessibles ou perdues ou encore celles n'ayant aucun ménage éligible.

Par ailleurs, si les ménages à l'intérieur d'une classe ont les poids du plan, alors R_j est calculé de manière plus appropriée comme le ratio entre le nombre pondéré de ménages interrogés et le nombre pondéré de ménages sélectionnés.

Au minimum, la situation géographique et d'autres informations utilisées pour la stratification des zones d'échantillon devraient être utilisées pour définir convenablement les classes de pondération pour l'ajustement de non-réponse dans les enquêtes sur les ménages. Par exemple, les zones d'échantillon au sein de chaque strate peuvent être disposées en fonction de la situation géographique ou d'autres critères que l'on juge associés aux variables de l'enquête et constitués en groupes d'environ 100 ménages. (Si les zones d'échantillon ont été sélectionnées à l'aide d'un échantillonnage systématique dans des listes ordonnées, alors le groupement suivant cet ordre sera généralement le choix approprié). Ensuite, on accorde le même poids de non-réponse à tous les ménages de l'échantillon qui se trouvent à l'intérieur du groupe (j), par proportion inverse au taux de réponse R_j du groupe.

D'autres classifications fondées sur les informations supplémentaires des types décrits ci-dessus pourraient si possible être créées. Quand d'autres variables (que ce soit au niveau de la zone ou du ménage) sont disponibles, chaque variable (ou chaque classification recoupée en deux ou plusieurs variables) peut être utilisée pour diviser tout l'échantillon en ensembles parallèles de classes de pondération. Chaque ensemble fournira la répartition, suivant la ou les variable(s) de classification concernée(s), du nombre de ménages sélectionné et du nombre interrogés avec succès sur la base desquels les taux de réponses peuvent être calculés dans chaque catégorie de classification.

7.2.5. Pondérations basées sur des sources externes plus fiables

Après l'ajustement des données de l'échantillon pour les probabilités d'échantillonnage et les taux de réponses différentiels, la distribution de l'échantillon selon le nombre et les caractéristiques des unités sera toujours différente des mêmes distributions disponibles issues de sources externes plus fiables comme le recensement de la population, les projections, les registres ou d'autres enquêtes à grande échelle. Normalement, on améliore la précision des estimations en pondérant davantage les données de l'échantillon de manière à accorder les distributions avec les informations externes.

L'application de cette étape ne nécessite pas la correspondance entre l'échantillon et la source externe au niveau de chaque ménage ou de chaque personne. Les ajustements de pondération sont réalisés sur la base de la comparaison des distributions d'échantillons et les distributions externes au niveau global.

> **Exemple:** Supposons que la distribution de l'échantillon selon une certaine caractéristique p_i se trouve dans la classe i de la classification interne de l'échantillon[49]. Soit P_i, la même proportion de source externe plus fiable. Alors les poids complémentaires à appliquer à toutes les unités de l'échantillon de cette classe seront simplement = (P_i/p_i). La distribution résultante pondérée de l'échantillon sera alors P_i.

[49] Les données ont peut-être été pondérées par les poids du plan et de non réponse, par exemple dans le calcul de cette proportion

Toutefois, un certain nombre de conditions devraient être examinées avant de prendre la décision d'utiliser les informations externes pour la pondération.

1. Il est d'abord nécessaire d'établir que l'on a besoin de ces pondérations, comme ce peut être le cas lorsque l'échantillon est petit ou lorsqu'il y a des erreurs évidentes – surtout des différences sérieuses avec l'échantillonnage probabiliste – dans le plan et la mise en œuvre. Ces considérations sont souvent importantes avec les enquêtes intensives sur les enfants travailleurs.

2. Une telle pondération doit être *pertinente* et *efficace* dans l'amélioration de la représentativité des résultats d'enquête.

3. Il est nécessaire que les informations externes soient clairement plus *fiables* que celles qui sont disponibles dans l'enquête elle-même.

4. Les informations externes devraient être cohérentes avec l'enquête, c'est-à-dire que les variables utilisées dans l'ajustement devraient être définies et mesurées de la même manière dans l'échantillon et la ou les source(s) externe(s). La cohérence est essentielle entre les sources lorsque plus d'une source externe sont impliquées.

5. Les informations externes devraient *couvrir diverses variables*, dans la mesure du possible.

Dans de nombreuses situations, il n'est pas suffisant de considérer la répartition à travers une seule caractéristique. Il est souhaitable de contrôler simultanément toutes les caractéristiques importantes. Toutefois, cela peut aboutir à l'application de trop de contrôles et, par conséquent, à de petites cellules d'ajustement et à de grandes variations dans les pondérations résultantes. Souvent, l'approche pratique consiste à contrôler simultanément un certain nombre de *distributions marginales,* plutôt que d'insister sur le contrôle d'une classification croisée trop détaillée de différentes variables.

Une méthode appropriée d'ajuster simultanément la distribution sur un certain nombre de contrôles externes est l'ajustement proportionnel itératif conventionnel ou méthode itérative du quotient proposée par Deming. L'idée principale consiste en la re-pondération de l'échantillon, dans le but d'accorder la distribution de l'échantillon avec la distribution externe, pour chaque variable de contrôle. Ce processuss est répété jusqu'à ce que toutes les variables impliquées soient étroitement accordées au même moment.

Dans les pays statistiquement moins développés, les données externes fiables et actualisées ne sont pas souvent disponibles. Il est donc nécessaire de se montrer prudent dans l'application de ces pondérations externes pour «corriger» les résultats de l'échantillon et éviter d'appliquer la correction à un niveau de classification trop détaillé. Un ajustement aveugle et complexe des résultats d'échantillon sur la base de données de qualité insuffissante a, malheureusement, été la pratique dans de nombreuses enquêtes.

7.2.6. Facteur d'extrapolation global

Ce terme désigne le facteur requis pour extrapoler les résultats de l'échantillon à l'ensemble de la population correspondante.

L'objectif de ce facteur consiste à isoler l'effet du taux d'échantillonage général (moyen), réduisant ainsi l'échelle de tous les autres facteurs impliqués dans la pondération, pour calculer la moyenne de certains nombres convenables tel que 1,0. En outre, la multiplication de tous les résultats de l'enquête par F compensera les effets bruts de toute sous-couverture, de toute non-réponse générale et de toute variation aléatoire sur la taille de l'échantillon réalisée.

Cette échelle n'affecte pas les estimations des proportions, des moyennes ou d'autres ratios de l'enquête, mais seulement les estimations des totaux ou des ensembles.

Toutefois, l'échelle appropriée est essentielle lorsque les résultats de différents domaines ou pays doivent être rassemblés. L'échelle des pondérations utilisée dans un tel regroupement, est telle que, pour chaque domaine, la somme des poids des cas d'échantillon est proportionnelle à la taille de la population du domaine.

En principe, différents facteurs d'extrapolation peuvent être pris en compte dans l'estimation des ensembles pour les différents types de quantités et de différents types d'unités comme les ménages et les personnes. Comme relevé ci-dessous, l'estimation des totaux (des ensembles de population) tirée des enquêtes par sondage nécessite normalement des informations contrôlées sur la taille et les ensembles pertinents de la population de base à l'aide de sources externes plus fiables.

7.3. Estimation des ratios et des totaux

7.3.1. Proportions, moyennes, ratios

Comme relevé au début de ce chapitre, calculer les facteurs de pondération lorsque cela s'avére nécessaire et les associer aux données pour chaque unité d'analyse de l'enquête, simplifie le processuss d'estimation des proportions, des moyennes, des ratios et même des statistiques plus complexes, en éliminant le besoin de faire appel explicitement à des complexités de base d'échantillonage. Le type d'estimateur le plus courant rencontré dans les enquêtes prend la forme d'un ratio de deux ensembles d'échantillon, à savoir *y* et *x*:

$$y = \sum_i y_i = \sum_i \sum_j w_{ij} \cdot y_{ij}$$

$$x = \sum_i x_i = \sum_i \sum_j w_{ij} \cdot x_{ij}$$

$$r = y/x$$

Le numérateur (y) et le dénominateur (x) peuvent être des variables explicatives – comme dans l'estimation du revenu moyen par tête d'une enquête ménages, où y est le revenu par tête et x le nombre total estimé de personnes de l'enquête. Pour chaque

ménage j de l'UEP i, y_{ij} renvoie à son revenu et x_{ij} à sa taille (=nombre de personnes, dans cet exemple). La quantité w_{ij} est le poids associé à l'unité.

Les moyennes ordinaires, les pourcentages et les proportions sont généralement des cas particuliers d'estimateur de rapport. Dans une moyenne, le dénominateur est une variable de dénombrement, c'est-à-dire que x_{ij} est identiquement égal à 1 pour tous les éléments de l'échantillon. Ce qui donne

$$\bar{y} = \frac{\sum_i \sum_j w_{ij} \cdot y_{ij}}{\sum_i \sum_j w_{ij}}$$

Pour une proportion (ou pourcentage), la condition complémentaire est que y_{ij} soit une dichotomie égale à 1 ou 0, selon que l'unité j possède ou non les caractéristiques dont la proportion est estimée. Par ailleurs, l'enquête peut aussi intégrer des statistiques plus complexes comme les différences, les sommes pondérées, les ratios ou d'autres fonctions des ratios. Ces données peuvent être estimées de façon analogue.

Il convient de relever un point très important. Une fois que les pondérations appropriées de l'échantillon ont été associées à chaque cas d'échantillon des données, l'estimation des statistiques comme celles qui précèdent ne nécessite aucune référence explicite à la structure de l'échantillon, tant que toutes les unités finales impliquées et leurs poids sont représentés. A titre d'exemple, dans l'équation ci-dessus, on se réfère à i, identificateur de l'UEP, simplement pour expliquer l'ensemble des unités incluses et non pour indiquer la dépendance de l'estimation sur toute caractéristique de la structure de l'échantillon autre que les pondérations individuelles.

Dans les échantillons stratifiés, la pratique courante consiste à utiliser des estimations de ratios «combinées», calculées à partir des résultats totaux à travers les strates pour assurer une plus grande stabilité (erreur quadratique moyenne plus faible):

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}}{\sum_h \sum_i \sum_j w_{hij} \cdot x_{hij}}$$

où, malgré les indices (dénotant l'élément j de l'UEP i dans la strate h), le numérateur et le dénominateur intègrent seulement les ensembles convenablement pondérés à travers les strates sur l'échantillon total (ou un domaine d'intérêt).

Dans un échantillon à plusieurs degrés, la probabilité de sélection d'une unité finale est le produit des probabilités aux diverses étapes de sélection. En estimant les proportions, les moyennes et d'autres types de ratios, comme précédemment, ce ne sont que les probabilités d'échantillonnage finales qui importent et non les détails aux diverses étapes. En fait, comme il a été relevé, hormis les poids, aucune autre complexité de l'échantillon n'apparaît dans cette estimation. Pour cette raison, les statistiques comme les ratios sont appelées «statistiques de premier ordre». Elles sont différentes des variances et d'autres «statistiques de second ordre», dont les procédures d'estimation prennent en compte la structure de l'échantillon et les poids des unités individuelles.

7.3.2. Estimation des totaux

Si les estimations des ratios de type ci-dessus sont souvent indiquées pour les moyennes et d'autres ratios, elles ont généralement besoin d'être modifiées lorsque l'objectif consiste à estimer les ensembles de population. Cela est particulièrement vrai pour des enquêtes ayant un plan à plusieurs degrés et un échantillon de petite taille. Ceci en raison du fait que, avec un plan de sélection d'échantillons à plusieurs degrés, la taille résultante de l'échantillon varie de façon aléatoire et, par conséquent, les ensembles estimés directement à partir de l'enquête peuvent comporter une erreur d'échantillonnage considérable. Le problème tend à être renforcé lorsque les estimations sont requises pour des sous-classes de population, dont la sélection n'est pas explicitement contrôlée dans le plan à plusieurs degrés.

Un autre problème crucial provient du fait que *les estimations d'ensembles sont directement biaisées proportionnellement à l'ampleur des erreurs de couverture*. En revanche, cela a un impact moins marqué sur les estimations des proportions, des moyennes, d'autres ratios et des statistiques plus analytiques. Généralement, les ensembles estimés à partir des enquêtes par sondage font l'objet de sous estimation. Ce fait se vérifie particulièrement pour les populations sur lesquelles il est difficile d'enquêter. Les enfants travailleurs constituent un exemple frappant.

La procédure appropriée pour l'estimation des ensembles de population est la suivante. Au lieu du facteur simple de la forme $\hat{Y} = F.y$, autrement dit l'extrapolation de l'ensemble de l'échantillon y par un facteur d'échelle constant F, l'inverse de la fraction d'échantillonnage générale, l'ensemble requis peut être exprimé sous la forme d'une estimation de ratio-type.

$$\hat{Y}_r = \frac{y}{x}.X$$

où y et x sont des totaux estimés à partir de l'échantillon; y étant la variable d'intérêt et x une variable auxiliaire pour laquelle un ensemble de population plus fiable X est disponible à partir d'une source externe.

> **Exemple:** Une enquête sur le travail des enfants a été utilisée pour estimer la *proportion* d'enfants engagés dans le travail. Le *nombre* d'enfants travailleurs est déterminé non pas directement à partir de l'estimation ci-dessus de l'enquête, mais en multipliant cette proportion par le nombre total d'enfants concernés dans la population, ce dernier étant estimé sur la base d'une source externe plus fiable comme les projections de population.

La valeur et l'applicabilité de cette procédure dépendent de plusieurs facteurs. Premièrement, le coefficient de corrélation entre y et x doit être positif et de préférence élevé, c'est-à-dire plus grand que 0,6 ou 0,7 au moins. Deuxièmement, X devrait être disponible avec plus de précision que l'estimation simple (x) de l'ensemble de population, qui peut être directement produite à partir de l'échantillon même. Troisièmement, X de la population et x de l'échantillon devraient être fondés sur des mesures essentiellement semblables sur la même population; une différence entre les deux introduirait un biais dans l'estimation.

7.3.3 Utilité de préparer des estimations non-biaisées simples

Il est néanmoins très important d'être capable de préparer des «estimations non-biaisées simples» de la forme $\hat{Y} = F.y$ à partir des données d'enquête, bien que celles-ci peuvent être par la suite affinées et modifiées en conséquence dans la production des estimations finales. Le terme «estimations non-biaisées simples» signifie que les estimations sont produites directement à partir des résultats d'enquête, sans recours aux données externes, en pondérant chaque observation par la proportion inverse à sa probabilité de sélection au sein de l'échantillon. Une telle estimation est produite pour la population totale, en pondérant simplement chaque valeur de l'échantillon par l'inverse de sa probabilité de sélection[50].

De telles estimations ne peuvent être préparées qu'avec l'échantillonnage probabiliste, c'est-à-dire pour les échantillons sélectionnés de sorte que chaque élément dans la population ait une probabilité connue et non nulle d'être sélectionnée. Pour préparer de bonnes estimations, il est également nécessaire que les problèmes de mise en œuvre de l'échantillonnage, notamment la non-réponse et la sous-couverture, ne déforment pas considérablement ces probabilités. De bonnes estimations simples impliquent également qu'un ajustement supplémentaire en vue d'améliorer leur précision n'est pas considérable. En conclusion, la capacité à produire de bonnes estimations non-biaisées simples indique que l'enquête a été efficacement conçue et mise en œuvre.

7.4. Commentaire sur l'estimation d'une petite zone

7.4.1. Le contexte

Il y a une demande croissante de statistiques, notamment les statistiques sur le travail des enfants, pour de petites zones, de petites sous-populations et d'autres domaines petits. Dans la pratique, il n'est pas possible – et non souhaitable de façon générale – d'étendre la taille de l'échantillon afin de remplir ces conditions à l'aide des estimations directes tirées de l'enquête. Il est nécessaire de développer des méthodes indirectes pouvant combiner différents types d'informations provenant de sources diverses pour produire des estimations plus fiables que celles que l'on obtiendrait sur la base de toute source utilisée individuellement.

Cette partie a pour objectif de fournir quelques notions d'introduction sur les techniques complexes et mises au point plus ou moins rapidement sur l'estimation d'une petite zone. La terminologie et la description utilisées dans les sous-sections suivantes sont en termes de combinaison conventionnelle tirée des données du recensement et des enquêtes par sondage. Un exemple serait utile afin de situer ces termes dans le contexte des enquêtes sur le travail des enfants. Dans ce contexte, nous considérons que le «recensement» est une enquête à grande échelle.

[50] Plus précisément, on écrit l'estimation $\hat{Y} = F \cdot y$ mentionné ci-dessus comme $\hat{Y} = \sum y_i/f_i$, où la somme de tous les éléments i dans l'échantillon et f_i est la valeur réelle de l'unité de la probabilité de sélection.

Supposons que dans une opération à grande échelle comme l'EMO, les estimations X_{igh} ont été obtenues pour le nombre d'enfants travailleurs par type d'activité (i), par groupe d'âge et de sexe (g) et par région administrative (h) du pays. Ces estimations sont fondées sur un ensemble de questions qui n'appliquent qu'une définition simplifiée du travail des enfants, mentionnée dans les enquêtes à grande échelle. Elles sont ainsi peu approfondies car centrées prioritairement sur des thèmes autres que celui du travail des enfants. Supposons que ces estimations donnent une idée réelle de la structure de variation (telle que par région), mais que l'ampleur réelle des estimations doit être ajustée. A supposer qu'une enquête plus approfondie, mais plus réduite, fournisse des estimations améliorées Y_{ig}, uniquement par type d'activité (i) et par groupe d'âge et de sexe (g); l'échantillon est trop petit pour permettre la répartition régionale (h).

Si l'on part de l'idée que pour chaque terme (i, g) donné, l'enquête étendue fournit une idée raisonnable de la variation régionale, nous pouvons estimer la variation par activité (i) et par région (h) de la manière suivante:

$$Y_{i.h} = \sum_{g} \left(\frac{X_{igh}}{X_{ig.}} \right) . Y_{ig.}$$

En principe, nous pouvons également estimer la répartition complète de la manière suivante:

$$Y_{igh} = \left(\frac{X_{igh}}{X_{ig.}} \right) . Y_{ig.}$$

mais généralement, une telle extension implique un degré d'incertitude plus élevé.

7.4.2. Estimations de petits domaines sur la base d'une utilisation combinée des données du recensement et de l'enquête

Le besoin des estimations pour de petits domaines

Le recensement peut fournir des données géographiquement détaillées, mais relativement peu fréquentes et moins actualisées, à l'inverse des enquêtes par sondage qui, bien que davantage fréquentes et actualisées, peuvent manquer de détails géographiques en raison des contraintes liées à la taille d'échantillon. Les données d'enquête par échantillonnage sont largemant utilisées pour déduire des estimations de totaux et de moyennes pour de vastes zones et domaines. Cependant, en dépit de grands progrès dans la capacité et la pratique de l'enquête, les estimateurs d'enquête «directs» les plus utilisés pour un petit domaine, fondés uniquement sur les unités du domaine peuvent causer des erreurs d'échantillonnage inacceptablement grandes en raison de la petitesse des tailles d'échantillon impliquées.

Toutefois, les demandes pour des statistiques opportunes et variées, pour des unités administratives de rang inférieur et de petites zones, se multiplient partout, non seulement dans les pays développés mais aussi dans les pays en développement. Ces besoins peuvent être comblés seulement en produisant des estimations plus fréquentes

et actualisées que ne le permet les recensements réalisés à grands intervalles, mais pouvant être classés avec beaucoup plus de détails que cela n'est possible dans les enquêtes par sondage de taille réduite. Les méthodes d'estimation de petites zones, destinées à fournir les renseignements actuels mais, détaillés ont été conçues à cette fin. L'idée fondamentale de ces procédures consiste à emprunter et à associer la force relative de plus d'une source dans la production de meilleures estimations (plus exactes et donc plus utiles).

Au cours de ces dernières décennies, des progrès majeurs et sophistiqués ont été réalisés dans la méthodologie des procédures d'estimation de petits domaines. L'objectif de cette section n'est pas de réviser ni d'évaluer ces procédures, mais aussi de fournir une introduction aux questions et aux approches de base[51].

Que sont les petits domaines?

Tout d'abord, il faudrait relever que le terme «domaine» ou «domaine d'estimation» est utilisé pour désigner la population ou une partie de la population pour laquelle des statistiques distinctes sont requises. Le terme «petite zone» ou «zone locale» est couramment utilisée pour désigner une zone géographique considérée comme «petite» dans un certain sens, telles que les provinces, les contés, les districts et d'autres circonscriptions administratives, les localités ou même les divisions du recensement ou les zones de dénombrement. Le terme domaine est plus général; il peut désigner les zones géographiques ou d'autres sous-population d'intérêt comme des groupes spécifiques par âge, par sexe, par nationalité, par race, ou d'autres caractéristiques.

Deuxièmement, il est important de relever que «petit» ne renvoie pas à la petitesse en termes de taille de la population du domaine d'intérêt, mais à celle de l'échantillon (nombre d'observations) disponibles en son sein. L'adjectif «petit» peut avoir des sens différents d'un contexte à un autre. Par exemple, dans les pays statistiquement plus développés où des vastes et fréquentes enquêtes par échantillonnage et/ou des sources administratives sont disponibles, il peut désigner des zones très locales ou de petits groupes de population. Dans de nombreux pays sous-développés, la production de statistiques utiles, même pour des provinces ou des districts vastes, peut nécessiter des techniques d'estimation de petits domaines en raison de l'impossibilité d'élargir suffisamment les tailles des enquêtes nationales. Les possibilités et l'adéquation de procédures d'estimation de petits domaines peuvent varier entre ces deux situations.

La classification des domaines suivant la catégorie de la taille peut être utile pour clarifier ces distinctions. En adaptant une classification proposée par Kish[52], nous pouvons distinguer:

- **Les principaux domaines**, entre 1 et 10 divisions de la population, par exemple les principales régions du pays, les principaux groupes professionnels, par secteur d'activité, par sexe et par âge, etc. Les enquêtes par échantillonnage sont souvent conçues pour fournir des estimations directes utiles pour ces divisions.

[51] Un texte complet sur ce sujet est Rao, J.N.K. (2003). *Small Area Estimation*. Wiley.

[52] Kish, L. (1980). Design and estimation for domains. *The Statistician*, 29(4) pp. 209-222.

- **Les domaines mineurs**, entre 10 et 100 divisions de la population, tels que chaque province en Indonésie et en Thaïlande ou chaque district au Kenya, classifiés en double entrée par profession et groupe de sexe et d'âge. *Dans de nombreux pays en développement, la priorité consiste à étendre les statistiques disponibles à ce type de niveau de détail.*
- **Les minis domaines**, entre 100 et 1 000 divisions de la population, tels que chaque comté aux Etats-Unis ou en Chine. A ce niveau de détail, la production des estimations fiables est souvent au-delà de la capacité de nombreux comtés, bien que certains pays plus développés sont capables de produire des estimations à ce niveau ou même aux niveaux de détail inférieurs.
- **Les domaines rares** peuvent être utilisés pour désigner environ 1 000 ou plus de sous-divisions, telles que des populations rares ou des classifications à multiples entrées.

7.4.3. Diversité de méthodes

Les premières méthodes d'estimation de petites zones mettaient l'accent sur les méthodes démographiques d'estimation de la population lors des années post-censitaires et sur l'estimation de la taille de la population active, des catégories de statut et d'autres caractéristiques d'intérêt. La plupart de ces méthodes utilisaient des données actuelles tirées de registres administratifs en rapport avec les données associées tirées du plus récent recensement de la population. Fondamentalement, ces méthodes d'«estimations symptomatiques» exploitent les rapports entre les variables symptomatiques (tels que le nombre de naissances, de décès, d'élèves inscrits, d'unités de logement etc. enregistrés localement et les variables d'intérêt à estimer (telle que la taille de la population locale). Généralement, de telles méthodes sont très spécifiques suivant la situation et dépendent de la qualité et du type de données administratives disponibles à cette fin.

Une seconde classe de méthodes appelée «estimation synthétique» emprunte les relations structurales entre les variables disponibles à un niveau détaillé (mais non actuelles) dans le recensement et impose ces relations d'une manière appropriée sur les données moins détaillées mais plus actuelles afin de produire des estimations à la fois détaillées et actuelles. La qualité des estimations résultantes dépend de la validité des relations importées.

Il peut aussi être utile d'associer les estimations indirectes aux estimations directes tirées de l'échantillon pour construire des «estimations composites» telle qu'une somme convenablement pondérée d'estimation directe et indirecte.

Les techniques sophistiquées plus récentes englobent les procédures telles que la technique «empirique de Bayes», la technique «hiérarchique de Bayes» et les «procédures empiriques de la meilleure prédiction linéaire non biaisée. Ces méthodes ont eu un impact considérable sur la pratique d'estimation de petites zones ces dernières années.

Certains aspects et formes des techniques d'estimation synthétiques mentionnées ci-dessus semblent à la fois faisables et pratiques pour une application dans des pays en

développement. L'idée essentielle qui sous-tend ce genre d'approche sera soulignée ci-dessous. Au-delà de cela, les procédures mentionnées ci-dessus ne sont pas réexaminées davantage. Au lieu de ceci, dans ce manuel d'échantillonnage pratique, il est utile de mettre l'accent sur la bonne approche de conception et d'utilisation de procédures d'estimation de petits domaines. L'adéquation de toute technique est propre à la situation et est déterminée par plusieurs critères:

- Disponibilité des données
- Exactitude des estimations
- Caractère pratique
- Acceptabilité parmi les utilisateurs

Voici résumées quelques leçons pratiques dans l'application de telles techniques.

> «De nombreuses leçons peuvent être tirées de la variété des méthodes disponibles. La première est que l'on peut trouver parmi elles une méthode meilleure que celle qu'on utilise arbitrairement à présent pour de petits domaines; il s'agit souvent de la méthode passive «nulle» qui consiste à se fier au dernier recensement décennal, qui peut être dépassé de 12 ans.
>
> Deuxièmement, il n'y a aucune méthode qui soit mieux dans toutes les situations. Des différences considérables existent entre les pays en ce qui concerne les sources et la qualité des données disponibles; la portée et la qualité de son recensement; l'ampleur, les contenus et les tailles des enquêtes par échantillonnage et surtout la portée et la qualité des registres administratifs. Cependant, les attitudes passives et négatives sont généralement injustifiées étant donné que chaque pays possède certaines ressources et que l'ingéniosité et les efforts permettent d'exploiter des ressources de données non utilisées. Ces données peuvent être d'origines apparemment différentes, mais potentiellement utiles en raison de la corrélation élevée des tailles de la population …
>
> Par ailleurs, le choix des sources et des méthodes devrait varier en fonction de la nature des statistiques, des estimations désirées et également des domaines auxquels ils appartiennent. Remarquez aussi les effets du temps écoulé depuis le dernier recensement… De façon plus générale, l'équilibre entre les biais d'un recensement et la variance d'une vaste enquête par échantillonnage penchera en faveur de cette dernière pendant les 10 ans qui séparent les recensements décennaux. La balance penchera aussi en faveur des registres moins fiables mais plus actuels. L'équilibre entre les enquêtes par sondage et les recensements ou les registres, devrait dépendre de la taille de l'échantillon, bien qu' une taille d'échantillon fixe comporte des avantages relatifs pour les populations plus réduites …
>
> Enfin, le choix entre les méthodes est plus difficile parce que la «meilleure» méthode n'est pas souvent évidente, même après l'évènement. Les erreurs d'estimation résultent principalement des biais et les «vraies valeurs» ne sont généralement pas disponibles pour mesurer directement ces biais. Les tests doivent être une combinaison de bases empiriques et de modèles, dépendant souvent au final (mais de maniére incertaine) de résultats tirés des recensements décennaux.

Les meilleurs méthodes et critères doivent être suivis à long terme dans une approche évolutive et avec patience.[53]»

«Nous… pouvons dessiner le simple tracé des progrès actuels et à venir. (1) Les méthodes sont utiles et utilisées maintenant pour les estimations post-censitaires des statistiques de la zone locale ... (2) Ces méthodes seront utilisées aussi pour d'autres statistiques et dans d'autres situations. (3) Les méthodes actuelles peuvent être et seront améliorées. (4) La force relative des méthodes différentes est difficile à prévoir et elle dépend des circonstances spécifiques; et peut être découvert lors d'essais empiriques spécifiques. (5) Le succès dépend d'abord de l'utilisation de meilleures données et ensuite de meilleures méthodes. … De bonnes sources de données sont les principaux moyens d'obtenir de meilleures statistiques… [nous devons] travailler dans le sens de la collecte d'autres données meilleures [et envisager également] des stratégies d'accumulation de données tirées des échantillons de petites zones.[54]»

7.4.4. Illustration d'une procédure

Nous fournissons ci-dessous une illustration simple d'une approche synthétique potentiellement utile. Comme simple illustration de la procédure, considérons qu'un échantillon fournit une estimation actualisée de la quantité ou du dénombrement (tels que les nombres par statut d'activité) y_g par groupe de sexe et d'âge (g) A partir du recensement, une répartition détaillée mais peu actualisée de la population Y_{gh} par groupe de sexe et d'âge (g) et par la petite zone (h) est connue. Cette dernière fournit une estimation de la répartition d'un groupe de sexe et d'âge par petite zone:

$$\left[Y_{gh}/Y_g\right]$$

En partant de l'hypothèse que cette répartition (la «relation structurale») est encore valable, les estimations de petites zones de la variable d'intérêt sont données en additionnant tous les groupes de sexe et d'âge g.

$$y_{.h} = \sum_g \left(Y_{gh}/Y_g\right).y_{g.}$$

L'idée fondamentale d'une forme plus générale de la procédure appelée «estimation de préservation de la structure» (SPREE) par ses auteurs[55] est illustrée dans le schéma ci-dessous.

A supposer que lors du recensement, la répartition détaillée mais peu actualisée su statut d'activité X_{igh} par groupe sexe-'âge (g) et par petite zone (h) est disponible. La répartition actualisée mais peu détaillée Y_{ig} du statut d'activité (i) par groupe d'âge (g) est disponible dans l'échantillon. Cela correspond à la répartition marginale de la

[53] Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domain). *International Statistical Review*, 48, 3-18.

[54] Kish, L. (1987). Discussion, in Small Area Statistics: An International Symposium, Platek et al (ed.). Wiley.

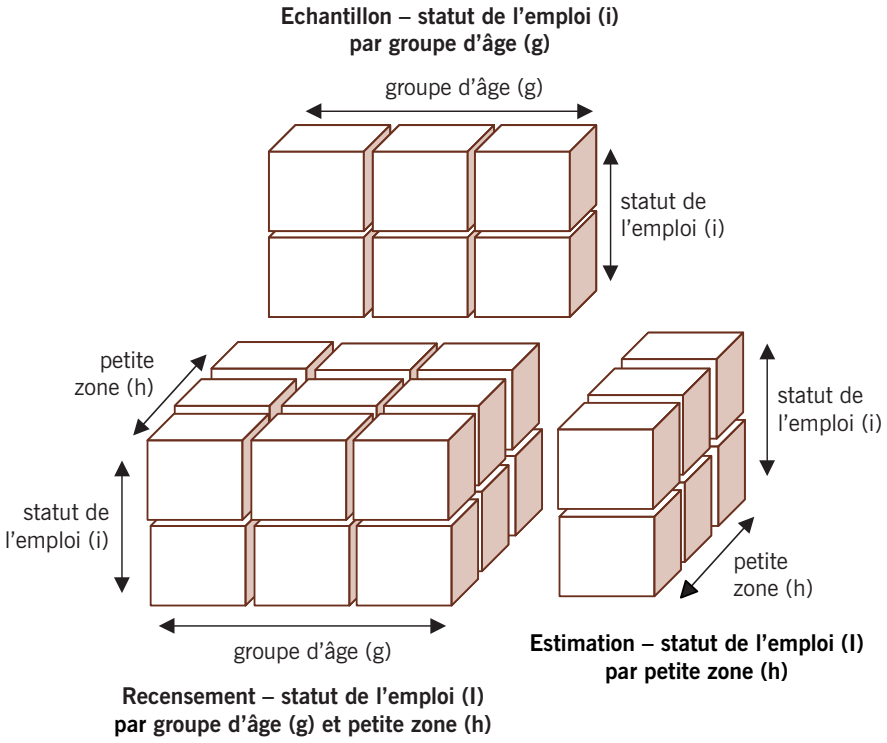
[55] Purcell, N.J., and Kish, L. (1980). Voir supra.

répartition complète du recensement, additionnée à travers les petites zones du pays. L'intérêt est d'estimer la répartition $Y_{i.h}$ de la situation d'activité (i) par petite zone (h), c'est-à-dire l'autre marge de la répartition complète du recensement, comme illustré sur le diagramme. L'estimation est donnée par

$$Y_{i.h} = \sum_{g} \left(\frac{X_{igh}}{X_{ig.}} \right) . Y_{ig.}$$

Elle est fondée sur l'hypothèse que le terme entre parenthèse, à savoir la distribution du statut du travail par groupe d'âge (i par g) à travers les petits domaines (h) est restée inchangée depuis le recensement. La distribution correspondante est «imposée» à chaque valeur (d'échantillon) actuelle par i et g, puis additionnée à travers tous les groupes d'âge pour obtenir des estimations sur le statut du travail de petites zones (bien sûr, sans classification par groupe d'âge).

Figure 2: Schéma d'illustration d'une procédure d'estimation de petite zone



Différentes formes de procédure peuvent être élaborées en fonction de la complétude des informations disponibles à partir du recensement et des types de contraintes des informations disponibles à partir de l'enquête, qui doit être satisfaite par les estimations résultantes de petites zones.

7.5. Importance des informations sur les erreurs d'échantillonnage

Les unités particulières qui viennent à être sélectionnées dans un échantillon particulier dépendent de la chance, les résultats possibles étant déterminés par les procédures précisées dans le plan de sondage. En d'autres termes, même si les informations éxigées sur chaque unité sélectionnée sont obtenues entièrement sans erreurs, les résultats tirés de l'échantillon font l'objet d'un degré d'incertitude, en raison de ces facteurs de chances affectant la sélection des unités. La variance d'échantillonnage (l'erreur d'échantillonnage, l'erreur type) est une mesure de cette incertitude.

Si les données de l'enquête font l'objet d'erreurs de diverses sources, les informations sur les erreurs d'échantillonnage sont d'une importance capitale pour une interprétation appropriée des résultats de l'enquête et pour une conception raisonnable des enquêtes par sondage. On ne peut pas trop mettre l'accent sur l'importance d'intégrer des informations sur les erreurs d'échantillonnage dans les rapports d'enquête.

Bien évidemment, l'erreur d'échantillonnage n'est qu'une composante de l'erreur totale des estimations de l'enquête et non la composante la plus importante. De même, la borne inférieure de l'erreur (et la plus facilement estimée) une enquête sera inutile si ce seul composant devient trop grand pour les résultats d'enquête pour ajouter des informations utiles avec toute la mesure de confiance par rapport à ce qui est déjà connu avant l'enquête.

En outre, les estimations de l'enquête sont typiquement requises non seulement pour l'ensemble de la population, mais aussi séparement pour de nombreux sous-groupes de la population. Généralement, l'ampleur relative de l'erreur d'échantillonnage par rapport à celle des autres types d'erreurs augmente à mesure que l'on passe des estimations de la population totale aux estimations de chaque sous-groupe et à la comparaison des sous-groupes. *Les informations sur l'ampleur des erreurs d'échantillonnage sont par conséquent essentielles dans la décision du degré de détail avec lequel les données d'enquête peuvent être totalisées et analysées de façon significative.*

Parallèlement, les informations sur les erreurs d'échantillonnage sont nécessaires pour le plan de sondage et l'évaluation. Si le plan est aussi déterminé par de nombreux autres facteurs (comme les coûts, la disponibilité des bases de sondage, le besoin de contrôler les erreurs de mesure), les décisions rationnelles sur le choix de la taille d'échantillon, l'allocation, la répartition en groupes, la stratification, les procédures d'estimation etc. ne peuvent se faire que sur la base d'une connaissance approfondie de leurs effets sur l'ampleur des erreurs d'échantillonnage des statistiques obtenues à partir de l'enquête.

Diverses méthodes pratiques et des logiciels ont été conçus pour calculer les erreurs d'échantillonnage et il n'y a aucune justification, dans la plupart des situations de continuer à omettre les informations sur les erreurs d'échantillonnage dans la présentation des résultats.

7.6. Procédures pratiques pour le calcul des erreurs d'échantillonnage

Les procédures pratiques d'estimation des erreurs d'échantillonnage doivent prendre en compte la structure réelle du plan, mais elles doivent être assez flexibles pour s'appliquer aux conceptions diverses du plan. Elles devraient être appropriées et convenables pour être appliquées à grande échelle (en produisant des résultats pour diverses statistiques et sous-classes) et économiques en terme d'effort et de coût impliqués.

Deux types de méthodes sont couramment utilisés, en pratique, pour les enquêtes fondées sur les plans complexes à plusieurs degrés:

1. Calcul à partir de comparaisons de certains ensembles pour «les sélections primaires» à l'intérieur de chaque strate de l'échantillon.

   Les erreurs d'échantillonnage peuvent aussi être calculées à partir de comparaisons des estimations des «reproductions» de l'échantillon. Nous pouvons distinguer deux différents types d'approche de cette classe;

2. Les estimations de la variance (et éventuellement des autres erreurs) fondées sur la comparaison de reproductions *indépendantes* de l'échantillon complet et

3. Les estimations de la variance fondées sur les comparaisons des reproductions se chevauchant et non indépendantes produites par le re-échantillonnage répété à partir de l'échantillon parent suivant un plan défini.

Le terme «sélection primaire» renvoie à l'ensemble des unités finales obtenues en appliquant une certaine procédure de sous-échantillonnage indiquée à chaque sélection d'une unité primaire d'échantillonnage. Les termes «grappe finale» et «reproduction» sont aussi utilisés pour décrire le même concept. Normalement, ce terme est simplement défini par l'ensemble d'unités finales provenant de la même UEP de l'échantillon – par exemple les unités sélectionnées dans le même ensemble d'un plan à deux degrés. Il y a des plans dans lesquels plus de deux sous-échantillons indépendants peuvant être sélectionnées à partir de la même UEP physique; dans ce cas, chacun de ces sous-échantillons constitue une sélection primaire.

Le terme «reproductions» renvoie aux parties de l'échantillon qui reflètent individuellement la structure (la répartition en grappes, la stratification, l'allocation, etc.) de l'échantillon total et diffère seulement de leur taille. Supposons par exemple, que nous sélectionnons 5 échantillons indépendamment de la même base de sondage, chaque échantillon ayant le même plan et la même taille. Rassemblons-les ensuite pour former notre échantillon final total. Chacune des cinq parties originales est une «reproduction» de l'échantillon total. Chacune d'elles a la même structure que l'échantillon total et ne diffère de lui qu'en taille d'échantillon. Dans cet exemple précis, les reproductions sont indépendantes.

7.6.1. Méthode 1. Estimation de la variance à partir de la comparaison entre les sélections primaires

Cette méthode est fondée sur la comparaison entre les estimations des sélections primaires indépendantes au sein de chaque strate d'un plan à plusieurs degrés. Voici

peut-être l'approche la plus simple de calculer les erreurs d'échantillonnage des statistiques ordinaires comme les proportions, les moyennes, les taux et d'autres ratios. Cette méthode peut facilement s'étendre à des fonctions plus complexes de ratios comme les différences ou les ratios, les ratios doubles, les indices, etc. Les équations de base se définissent comme suit.

Soit une population totale Y obtenue en additionnant les valeurs individuelles Y_{hij} pour les éléments j sur l'UEP i et ensuite toute les UEP et les strates h de la population:

$$Y = \sum_h Y_h = \sum_h \sum_i Y_{hi} = \sum_h \sum_i \sum_j Y_{hij}$$

L'expression ci-dessus est évaluée en additionnant les valeurs pondérées appropriées sur les unités de l'échantillon:

$$y = \sum_h y_h = \sum_h \sum_i y_{hi} = \sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}$$

Pour l'estimateur de ratio combiné des deux ensembles y et x

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} \cdot y_{hij}}{\sum_h \sum_i \sum_j w_{hij} \cdot x_{hij}}$$

l'expression générale de la variance est

$$\mathrm{var}(r) = \sum_h \left[(1 - f_h) . \frac{a_h}{a_h - 1} . \sum_i (z_{hi} - \frac{z_h}{a_h})^2 \right]$$

où a_h est le nombre de sélections primaires de la strate h, f_h le taux d'échantillonnage dans la strate et la variable de calcul z définie comme

$$z_{hi} = \frac{1}{x} . (y_{hi} - r . x_{hi}); \quad z_h = \sum_i z_{hi}$$

Cette approche est fondée sur les hypothèses suivantes du plan de l'échantillon.

1. La sélection de l'échantillon est indépendante entre les strates.
2. Deux ou plusieurs sélections primaires sont tirées de chaque strate (a_h>1).
3. Ces sélections primaires sont tirées de façon aléatoire, indépendamment et avec remise.
4. Le nombre de sélections primaires est assez grand pour une utilisation valable de l'estimateur du ratio et l'approximation prise en compte dans l'expression de sa variance.
5. Les quantités x_{hi} du dénominateur (qui correspondent souvent aux tailles des échantillons par UEP) sont raisonnablement uniformes en taille au sein des strates[56].

[56] La condition qui vient d'être mentionnée concerne le fait de garder le biais de l'estimateur du ratio réduit. En pratique, la condition est que de façon idéale, la variance relative, var(x)/x, devrait être inférieure à 0,1; et ne devrait en aucun cas dépasser 0,2 lorsqu'une estimation de ratio est utilisée.

Les formules ci-dessus, d'estimation de la variance sont très simples, en dépit de la complexité du plan, car elles sont fondées uniquement sur les sommations pondérées des sélections primaires et de l'identification des strates.

Un point des plus importants est que la complexité de l'échantillonnage dans les UEP ne semble pas compliquer de façon explicite la procédure d'estimation. Aucun calcul séparé des composantes de la variance n'est requis. Cela confère une grande flexibilité dans le traitement des divers plans d'échantillonage. Cette flexibilité est l'un de ses principaux atouts et la raison de son utilisation répandue dans les activités d'enquête.

De plus, cette méthode nécessite l'élaboration de différentes formules d'estimation de la variance pour différents types de statistiques, difficilement applicable aux statistiques très complexes. Dans le rapport descriptif des résultats d'enquête, la plupart des statistiques d'intérêt sont sous forme de proportions, de moyennes, de taux, de ratios ou d'une des fonctions de ces données. Pour ces statistiques, les formules de base ci-dessus s'appliquent sur l'estimation de la variance, ce qui rend la procédure plus directe. D'autres mesures plus complexes peuvent être utilisées dans le rapport analytique des résultats d'enquête. Le calcul de la variance peut être plus difficile pour des mesures qui utilisent la méthode décrite ci-dessus.

7.6.2. Méthode 2. Variance tirée des reproductions indépendantes

Il s'agit en principe d'une procédure très simple capable de gérer les estimations, quelle que soit leur complexité. Elle a été largement utilisée pour les enquêtes officielles en Inde et ailleurs. La méthode est fondée sur l'idée que l'échantillon parent peut être considéré comme formé de reproductions indépendantes ou de sous-échantillons, chacun reflétant toute la complexité de l'échantillon parent, ne lui étant différent que par sa taille. Avec ces hypothèses, les estimations reproduites peuvent être considérées comme indépendantes et les variables aléatoires indépendantes et identiquement distribuées (IID), de telle sorte que la variabilité entre elles leur confère, sous une forme très simple, une mesure de la variance de l'estimateur d'échantillon global.

La faiblesse de cette méthode est que dans de nombreuses situations, l'échantillon total ne peut être divisé en nombre suffisant de reproductions indépendantes de taille adéquate pour que la méthode soit applicable; sa force réside dans sa simplicité quand elle est applicable.

La procédure

La condition élémentaire de la méthode est que, par plan d'échantillonnage ou à la suite de la sélection d'échantillon, il devrait être possible de diviser l'échantillon parent en reproductions plus ou moins *indépendantes*, chacune possédant foncièrement le même plan que l'échantillon parent. Dans un plan à plusieurs degrés par exemple, l'échantillon parent doit être divisé au niveau des UEP, c'est-à-dire divisé de manière exhaustive en un certain nombre de reproductions ne se chevauchant pas, chacune étant constituée d'un ensemble distinct indépendamment sélectionné (idéalement et indépendamment dénombré) d'UEP. Pour une estimation combinée sur un certain nombre de strates, chaque reproduction doit elle-même être un échantillon stratifié couvrant toutes les strates.

Avec des reproductions indépendantes, chacune fournissant une estimation valable du même paramètre de la population d'intérêt, les résultats de la théorie «d'estimateur de la variance reproduite indépendante» peuvent être directement appliqués. La variance de la moyenne simple de n estimations reproduites y_j

$$\bar{y} = \frac{\Sigma_j y_j}{n}\ ; \quad \mathrm{var}(\bar{y}) = \frac{1}{n}.\left[\frac{\Sigma_j \left(y_j - \bar{y}\right)^2}{(n-1)}\right]$$

fournit une estimation de la variance de la même estimation tirée de l'échantillon total, exactement pour une estimation linéaire y ou approximativement pour une estimation non linéaire $\tilde{y}$. Une sorte d'estimateur conservateur (donnant une valeur plus élevée) est obtenue en remplaçant $\bar{y}$ dans l'addition par $\tilde{y}$; c'est-à-dire en écrivant:

$$\mathrm{var}(\tilde{y}) = \frac{1}{n}.\left[\frac{\Sigma_j \left(y_j - \tilde{y}\right)^2}{(n-1)}\right]$$

L'expression ci-dessus peut-être modifiée pour incorporer la correction de la population finie, si cela est important.

Construction de reproductions indépendantes

Il est utile de commencer par définir les conditions idéales d'application de la procédure. La condition de base est que l'échantillon parent (total) soit constitué d'un certain nombre de sous-échantillons ou reproductions indépendantes, ayant chacun le même plan et les mêmes procédures, mais sélectionnés et mis en œuvre indépendamment.

La condition de procédures courantes et indépendantes dans la construction des reproductions tirées de l'échantillon s'applique à la sélection de l'échantillon ainsi qu'à la collecte et à l'estimation de données.

Sélection de l'échantillon. Les reproductions devraient être conçues selon le même plan d'échantillon, sur la base de la même base de sondage et du même type d'unité, de système de stratification, de degrés d'échantillonnage et de méthode de sélection, etc. que l'échantillon parent. En extrayant de nombreuses reproductions de la même population, l'indépendance exige que chaque reproduction soit placée dans la base de sondage avant que la suivante soit tirée et la procédure de sélection aléatoire est appliquée séparément à chaque sélection.

Collecte de données. A la suite de la sélection de l'échantillon, la procédure de collecte de données doit être la même et appliquée indépendamment de chaque reproduction. La collecte de données désigne diverses étapes de l'ensemble du processus de mesure, y compris la conception du questionnaire, le recrutement et la formation du personnel, le mode et les procédures de collecte de données, l'organisation du travail de terrain, la supervision et le contrôle, l'enregistrement et le codage des réponses, la saisie des données, etc.. L'application indépendante des procédures courantes de collecte de données nécessite par exemple que les ensembles indépendants du personnel de terrain

(superviseurs, intervieweurs, codeurs, etc.) tirés en principe du même groupe soient utilisés pour diverses reproductions.

Estimation. Une procédure d'estimation courante désigne non seulement la forme mathématique de l'estimateur particulier utilisé, mais également toutes les autres étapes impliquées dans le calcul des estimations finales tirées des données de l'enquête – notamment l'édition des données, l'imputation, le traitement des observations aberrantes, la pondération et d'autres ajustements. L'application indépendante signifie que toutes les étapes de la procédure d'estimation ont été appliquées séparément à chaque reproduction. Par exemple, si les données de l'échantillon sont pondérées pour s'harmoniser avec certains totaux de contrôle de la population, cela sous-entend que les poids pertinents sont déterminés indépendamment de chaque sous-échantillon reproduit, distinct de l'utilisation d'un ensemble commun de poids déterminés sur la base d'un ensemble de données complet.

De brefs commentaires sur divers aspects de l'application de la procédure dans la pratique suivent.

Approximation des reproductions indépendantes

Dans la pratique, les conditions ci-dessus sont rarement remplies exactement. Par exemple, si la procédure d'estimation est complexe, la reprise de toutes les étapes pour chaque reproduction peut se révéler trop coûteuse et chronophage. Par conséquent, si des estimations distinctes sont produites pour chaque reproduction (comme il se doit pour la procédure d'estimation de la variance à appliquer), certaines étapes de la procédure – notamment l'imputation, la pondération et l'ajustement des résultats par rapport aux totaux de contrôle externe – s'appliquent une seule fois à l'échantillon dans son ensemble. Les résultats peuvent être différents si ces étapes sont appliquées séparément et indépendamment des résultats de l'échantillon de chaque reproduction. L'indépendance stricte des procédures de collecte de données est encore plus difficile à mettre en œuvre. Cela nécessiterait l'organisation et la mise en oeuvre de nombreuses étapes dans le processus de mesure indépendamment de chaque reproduction, impliquant éventuellement une forte augmentation du coût et des inconvénients. (Une telle séparation sous une forme indiquée peut bien sûr être utile dans l'évaluation des erreurs de non-échantillonnage).

La condition la plus déterminante est peut-être celle de la sélection indépendante des reproductions suivant le même plan. Théoriquement, l'échantillon total peut être formé *en associant* les sous-échantillons indépendants. Cependant, c'est souvent un problème de *partitionnement* d'un échantillon existant en un nombre plus ou moins de sous-échantillons. Il est important de relever que dans un plan à plusieurs degrés, le partitionnement de l'échantillon devrait se faire au niveau de la sélection primaire, c'est-à-dire que tous les éléments de l'échantillon à l'intérieur d'un SP devraient être associés à la même reproduction. Pour estimer la variance totale à travers les strates d'un échantillon stratifié, chaque reproduction doit elle-même être un échantillon stratifié parallèle à l'échantillon parent.

L'échantillon peut être divisé en reproductions au moment de la sélection ou à la suite, après la sélection. Soit un échantillon systématique dans lequel les unités primaires doivent être sélectionnées par intervalle I; l'échantillon peut être sélectionné sous la forme de n reproductions, sélectionnées chacune systématiquement avec une origine choisie au hasard et un intervalle de sélection = (n.I). Une solution de rechange plus indiquée et courante consiste à sélectionner l'échantillon total une seule fois, mais de telle sorte qu'il puisse être divisé par la suite en sous-échantillons indépendants de façon générale et reflétant le plan de l'échantillon total. Soit par exemple un échantillon systématique de 500 UEP devant être divisées en 20 reproductions ayant chacune une taille de 500/20 = 25 UEP. On peut imaginer une liste ordonnée de 500 unités d'échantillon comme divisées en 25 «zones», comprenant chacune 20 unités adjacentes. Une reproduction sera composée d'une unité prise dans chaque zone: par exemple, la première unité de chaque zone formant la première reproduction, la seconde unité de chaque zone formant la seconde reproduction. En effet, un tel plan simple peut être appliqué avec une grande flexibilité et permet de nombreuses variations directes. Les unités de l'échantillon total peuvent être sélectionnées avec des probabilités uniformes ou variables: le plan de sous-échantillonnage ci-dessus conserve les probabilités relatives de sélection. Si l'échantillon original est stratifié, on peut ordonner la strate des unités sélectionnées par strate et diviser la liste entière en zones égales pour l'application de la procédure ci-dessus. L'effet de la stratification originale sera reflété dans les reproductions si le nombre d'unités devant être sélectionnées est assez grand pour que la totalité ou la majorité des strates soit représentée dans chaque reproduction. Subsidiairement, les unités peuvent être recoupées par zone et par strate, c'est-à-dire que chaque strate est divisée en un certain nombre de zones et chaque zone reliant les unités (les UEP de l'échantillon) à travers un certain nombre de strates. Deming (1960) fournit de nombreux exemples et extensions de telles procédures[57].

Choix du nombre de reproductions

Dans la plupart de plans à plusieurs degrés, le nombre de sélections primaires impliqué est réduit, ce qui limite le nombre de reproductions dans lesquelles l'échantillon peut être divisé. Il y a bien sûr des échantillons dans lesquels le nombre total de SP disponibles est si inadapté que le nombre de reproductions et le nombre d'unités par reproduction doivent plutôt être faibles. Dans cette situation, la méthode de reproduction indépendante est inappropriée pour l'estimation de la variance. Toutefois, lorsque le plan de l'échantillon le permet, le choix doit encore se faire entre les deux extrêmes: avoir de nombreuses petites reproductions ou bien n'en avoir que peu, mais de grandes tailles. Si de nombreuses reproductions sont créées, le nombre de SP par reproduction peut devenir trop faible pour refléter la structure de l'échantillon total. L'estimation de la variance en sera généralement biaisée. D'un autre côté, les variances estimées à partir de petit nombre de reproductions seulement tendent à être instables. En général, il n'y a aucun consensus en ce qui concerne le choix le plus approprié. Kish (1965, Section 4.4), par exemple, résume la situation de la manière suivante:

> «Mahalanobis et Lahiri ont fréquemment employé 4 reproductions ... Tukey et Deming ont souvent utilisé 10 reproductions ... Jones présente les raisons et les

[57] Deming, W.E. (1960). *Sample Design in Business Research*. Wiley.

règles d'utilisation de 25 à 50 reproductions. Généralement, je privilégie également un grand nombre allant peut être de 20 à 100[58].»

Le premier argument en faveur du fait d'avoir de nombreuses reproductions (chacune étant nécessairement constituée d'un nombre réduit d'unités) est que l'estimateur de la variance est plus précis et les statistiques $\overline{y}$ (moyenne des estimations reproduites) suivent presque une loi normale. La précision de l'estimateur de variance baisse à mesure que le nombre de reproductions est réduit. Par ailleurs, pour une valeur donnée de la variance ou de l'erreur type, l'intervalle associé à un niveau de confiance donné devient plus large. Une autre considération est qu'avec un petit nombre de reproductions, il est nécessaire de supposer que les estimations reproduites individuelles y_j sont normalement réparties, bien que de légères divergences par rapport à la normalité ne sont généralement pas signifiantes; heureusement, l'hypothèse de la normalité de y_j est améliorée à mesure que le nombre de sélections primaires par reproduction augmente. Dans tous les cas, lorsque le nombre de reproductions est considérable, il est nécessaire de supposer seulement que la moyenne $\overline{y}$ est normalement répartie.

Par ailleurs, le fait d'avoir très peu de reproductions, mais d'une grande taille comporte aussi des avantages statistiques.

1. Avec une grande taille d'échantillon par reproduction, les estimations reproduites individuelles y_j sont plus stables et presque plus normalement réparties. Cela est utile à l'inférence statistique.

2. Les estimations reproduites y_j, et plus encore leur moyenne $\overline{y}$ sont également plus proches de l'estimation $\widetilde{y}$ fondée sur l'échantillon total des statistiques non linéaires. Cela facilite l'extension de la méthode d'estimation de la variance aux statistiques non linéaires, ce qui constitue la principale justification de son usage.

3. Plus important encore, le fait d'augmenter le nombre d'unités primaires par reproduction permet plus facilement de refléter la structure de l'échantillon total dans chaque reproduction, ce qui réduit le biais dans l'estimateur de variance.

   Il convient également de relever certaines considérations pratiques et des objectifs plus vastes pour adopter un plus petit nombre de reproductions, chacune étant par conséquent plus large et potentiellement plus complexe dans le plan:

4. Avec très peu de reproductions, il y a moins de perturbations du plan général causées par le besoin de sélectionner l'échantillon sous forme de reproductions indépendantes.

5. Le coût supplémentaire et la difficulté de la mesure et de l'estimation séparées sont plus faibles.

6. Avec un nombre plus petit de reproductions, il est plus fiable de randomiser convenablement l'allocation du travail des intervieweurs, des codeurs, etc. pour mesurer les composantes de la variance non liées à l'échantillonnage.

[58] Kish, L. (1965). *Survey Sampling*. Wiley.

7. Les plans reproduits ou «en réseau» peuvent être utiles pour une vérification plus générale des procédures et des résultats d'enquête. Ces objectifs sont mieux réalisés lorsque le nombre de reproductions à traiter est faible.

8. Il en va de même de l'affichage séparé des résultats d'enquêtes par reproduction pour donner à l'utilisateur une impression claire de la variabilité des résultats de l'enquête par échantillonnage.

Il est bien sûr possible d'avoir un plus grand nombre de reproductions pour une estimation plus stable de l'erreur type et de les diviser en un plus petit nombre d'objectifs (6), (7) et (8).

Face aux considérations et aux opinions conflictuelles ci-dessus, il n'est pas possible de formuler des recommandations sur le choix indiqué du nombre et de la taille des reproductions. Avec environ 100 à 1 000 UEP dans l'échantillon, une règle simple, jugée plutôt raisonnable, consiste à commencer par égaliser le nombre de reproductions et le nombre d'UEP par reproduction à la racine carrée du nombre total donné d'UEP de l'échantillon. Par exemple, avec en moyenne plus de 200 UEP d'échantillon, ce qui précède aboutira à environ 15 reproductions ayant chacune 15 UEP. De façon similaire, avec en moyenne plus de 600 UEP, on pourrait commencer par examiner environ 25 reproductions ayant chacune 25 UEP.

7.6.3. Méthode 3: Comparaison des reproductions de l'échantillon total sur la base du re-échantillonnage répété

Ces procédures sont plus complexes et à grande capacité de calcul, mais elles peuvent s'appliquer aux statistiques quelle que soit leur complexité. L'idée de base est celle du «re-échantillonnage répété». Cette approche désigne la classe de procédures pour le calcul des erreurs d'échantillonnage des plans et des statistiques complexes dans lesquels les reproductions à comparer sont produites par le re-échantillonnage répété du même échantillon parent. *Chaque reproduction est conçue pour refléter la complexité totale de l'échantillon parent.*

Contrairement à la procédure décrite au paragraphe 7.6.2, les reproductions en elles-mêmes ne sont pas indépendantes (en fait, elles se chevauchent) et des procédures spéciales sont requises pour contrôler le biais dans les estimations de la variance générées par comparaison de telles reproductions.

Comparées aux méthodes du paragraphe 7.5.1, les méthodes de re-échantillonnage répété ont pour inconvénient une plus grande complexité et un travail de calcul. Elles tendent aussi à être moins flexibles dans les plans d'échantillons traités.

Toutefois, elles ont l'avantage de ne pas exiger une expression explicite de la variance de chaque statistique particulière. La même expression de base (donnée ci-dessous) s'applique à chaque statistique.

Dès que les reproductions ont été créées, aucune autre référence explicite n'est requise, même à la structure de l'échantillon.

Les procédures sont aussi plus englobantes: en répétant toute la procédure d'estimation indépendamment de chaque reproduction, l'effet des diverses complexités, notamment chaque étape d'une procédure de pondération complexe, peut être intégrée aux estimations de la variance produites.

La condition technique importante est la construction d'un ensemble approprié de reproductions aux termes desquelles la variance requise peut être estimée comme expliqué ci-dessous.

Les différentes procédures de re-échantillonnage disponibles diffèrent quant à la manière dont les reproductions sont générées à partir de l'échantillon parent et des formules d'estimation de la variance correspondantes évoquées. Il y a trois procédures générales connues sous les noms de Reproduction Répétée du Jacknife (JRR), de Répétition Répétée Equilibrée (BRR) et de «bootstrap».

Peut-être que la JRR fournit généralement la méthode la plus polyvalente et la plus indiquée. Elle est définie ci-dessous.

La Reproduction Répétée du Jacknife (JRR)

La Reproduction Répétée du Jacknife (JRR) est une classe de méthodes d'estimation des erreurs d'échantillonnage, à partir des comparaisons des reproductions de l'échantillon générées à travers l'échantillonnage répété d'un même échantillon parent. Chaque reproduction doit elle-même être un échantillon représentatif et refléter toute la complexité de l'échantillon parent.Toutefois, lorsque les reproductions ne sont pas indépendantes, des procédures spéciales sont nécessaires à leur construction pour éviter le biais dans les estimations résultantes de la variance.

Présentée au départ comme une technique de réduction du biais, la méthode du Jacknife a été largement expérimentée et utilisée à ce jour pour l'estimation de la variance. Nous préférons la JRR à d'autres méthodes comme la Reproduction Répétée Equilibrée, car la JRR est généralement plus simple et plus flexible.

Le modèle de base d'application du JRR dans le contexte décrit ci-dessus peut être résumé de la manière suivante. Soit un plan dans lequel deux ou plusieurs unités primaires ont été sélectionnées indépendamment de chaque strate de la population. Au sein de chaque unité de sélection primaire (UEP), on peut inclure le sous-échantillonnage, quelle que soit la complexité, y compris la pondération des unités finales.

Dans la version «standard», chaque reproduction du JRR peut être formée en abandonnant une UEP dans une strate précise et en augmentant convenablement le poids des UEP correspondantes dans cette strate pour obtenir une estimation alternative mais valable de celle obtenue dans l'échantillon total.

Ce qui précède implique la création d'autant de reproductions que le nombre d'unités primaires de l'échantillon. Le travail de calcul impliqué est parfois réduit en limitant le nombre de reproductions requis. Par exemple, les UEP peuvent être groupées au sein des strates et former les reproductions du JRR en éliminant un groupe à la fois. Cela n'est possible que quand chaque strate renferme plusieurs unités. Subsidiairement, ou en outre, les groupements des unités peuvent être coupés à travers les strates.

Troisièmement, il est possible de définir les reproductions d'une manière standard («méthode du Jacknife consistant à supprimer une UEP à la fois»), mais construire et utiliser un seul de ces sous-ensembles.

Lorsque l'échantillon ou une de ses parties est un échantillon direct des unités finales ou de petites grappes, de sorte que le nombre de reproductions sous le JRR «standard» est trop élevé pour être pratique, un certain groupement des unités est inévitable. Normalement, la procédure approppriée pour réduire ce nombre serait de former de nouvelles unités de calcul en constituant des groupements aléatoires d'unités au sein de la strate.

En dehors de ce qui précède, il y a peu de besoin ou de motivation d'introduire les types de raccourcis ci-dessus ou d'approximations de la méthode standard du JRR dans les enquêtes sociales réelles utilisant des plans d'échantillon à plusieurs degrés.

En résumé, la méthode standard du JRR implique ce qui suit.

Base théorique

Premièrement, la base théorique se définit de la manière suivante. Elle implique la généralisation à partir des statistiques linéaires comme un simple ensemble par rapport aux statistiques, quelle que soit leur complexité.

Soit une reproduction formée en introduisant une UEP donnée i dans une strate h et en augmentant convenablement le poids des UEP restantes (a_h-1) dans cette strate pour compenser l'UEP manquante. L'estimation d'un simple ensemble (total) de cette reproduction est

$$y_{(hi)} = \sum_{k \neq h} y_k + \frac{a_h}{a_h - 1}.(y_h - y_{hi}) = y - \frac{a_h}{a_h - 1}.(y_{hi} - \frac{y_h}{a_h})$$

Avec la moyenne de ces estimations de la strate, $y_{(h)} = \sum_i y_{(hi)} / a_h$

et la moyenne sur toutes les reproductions $a = \sum ah$, $\tilde{y} = \dfrac{\sum_h \sum_i y_{(hi)}}{\sum_h a_h}$

L'expression de la variance de toute statistique peut s'écrire de diverses formes statistiquement équivalentes.

$$var_1(y) = \sum_h [(1 - f_h).\tfrac{a_h - 1}{a_h}.\sum_i (y_{(hi)} - y_{(h)})^2]$$

$$var_2(y) = \sum_h [(1 - f_h).\tfrac{a_h - 1}{a_h}.\sum_i (y_{(hi)} - \tilde{y})^2]$$

$$var_3(y) = \sum_h [(1 - f_h).\tfrac{a_h - 1}{a_h}.\sum_i (y_{(hi)} - y)^2]$$

La forme standard de la variance de l'ensemble y simple

$$var(y) = \sum_h \left[(1 - f_h).\frac{a_h}{a_h - 1}.\sum_i (y_{hi} - \frac{y_h}{a_h})^2 \right]$$

est remplacée dans la méthode du JRR par l'une des expressions ci-dessus (généralement la dernière des trois, var_3, étant donné qu'elle est plus conservatrice). Etant basée presque sur la totalité de l'échantillon, les estimations comme $y_{(hi)}$, $y_{(h)}$ et même plus, leur moyenne totale $\tilde{y}$ sont censées se rapprocher de l'estimation de l'échantillon y, même pour les statistiques complexes. Il en résulte que leur variance, exprimée par l'une des trois formes fournit aussi une mesure de la variance de y. *Cela s'applique aux statistiques y, quelle que soit sa complexité,* pas seulement à l'ensemble simple. En revanche, la forme standard var(y) s'applique seulement quand y est un ensemble simple. Dans ce cas, toutes les formes ci-dessus sont en fait équivalentes.

Application aux statistiques complexes

Soit z, une estimation de l'échantillon total, quelle que soit sa complexité, et $z_{(hi)}$, l'estimation produite à l'aide de la même procédure après l'élimination de l'unité primaire i de la strate h et en augmentant le poids des unités restantes (a_h-1) de la strate par un facteur approprié g_h (voir ci-dessous). Soit $z_{(h)}$, la moyenne simple de $z_{(hi)}$ sur les unités de l'échantillon a_h de h. La variance de z est alors estimée comme:

$$\mathrm{var}(z) = \Sigma_h \left[(1 - f_h) . \frac{a_h - 1}{a_h} . \Sigma_i \left(z_{(hi)} - z_{(h)} \right)^2 \right] \qquad (1)$$

Le principal avantage d'une procédure comme la JRR est que, dans les conditions assez générales d'application de cette procédure, la même formule relativement simple d'estimation de la variance peut s'appliquer à z, quelle que soit la complexité.

Normalement, le facteur g_h est défini par (2.a)

$$g_h = a_h / (a_h - 1) \qquad (2.a)$$

mais pour des raisons relevées ci-dessous, il est préférable d'utiliser la version (2.w):

$$g_h = w_h / (w_h - w_{hi}) \qquad (2.b)$$

où , $w_h = \Sigma_i w_{hi}, w_{hi} = \Sigma_j w_{hij}$, la somme des poids des unités finales j dans la sélection primaire i.

La forme (2.b) est proposée parce qu'elle conserve le poids total des cas de l'échantillon inclus sans changement à travers les reproductions créées. Avec des poids de l'échantillon à une échelle telle que leur somme soit égale ou proportionnelle à un total externe plus fiable de la population, les ensembles de population de l'échantillon peuvent être estimés plus efficacement, souvent avec la même précision que les proportions et les moyennes.

Une autre variation possible, qui peut être mentionnée, consiste à remplacer $z_{(h)}$, la moyenne simple de $z_{(hi)}$ sur les unités de l'échantillon a_h de h, par l'estimation de l'échantillon total z_h de la strate h:

Soit $z_{(h)}$ la moyenne simple de $z_{(hi)}$ sur les unités de l'échantillon a_h de h. La variance de z est alors estimée comme:

$$\mathrm{var}(z) = \Sigma_h \left[(1 - f_h) . \frac{a_h - 1}{a_h} . \Sigma_i \left(z_{(hi)} - z_h \right)^2 \right] \quad (3)$$

La version (1) tend à fournir une estimation «conservatrice» de la variance, mais normalement la différence reste mince par rapport à (3).

Les estimations de la variance du JRR prennent en compte l'effet sur la variance des aspects du processus d'estimation qui sont autorisés à varier d'une reproduction à l'autre. En principe, cela peut englober les effets complexes comme l'imputation et la pondération. Très souvent il n'est pas possible en pratique de renouveller entièrement de telles opérations pour chaque reproduction.

7.7. Application des méthodes dans la pratique

Bien que les hypothèses de bases relatives à la structure de l'échantillon pour l'application de la méthode soient raisonnablement bonnes dans de nombreuses enquêtes à grande échelle auprès des ménages, ces hypothèses ne sont pas souvent exactes. Les solutions pratiques peuvent néanmoins être trouvées dans la plupart des situations. En voici les plus courantes.

Problème	Solution pratique générale
L'échantillonnage systématique des unités primaires, une procédure générale et commode, ne donne pas de façon précise un minimum de deux sélections primaires indépendantes par strate.	Une pratique habituelle consiste à coupler les unités adjacentes pour former des strates à utiliser dans les calculs. On suppose que les unités appareillées ont été sélectionnées indépendamment au sein de la strate ainsi modifiée.
Parallèlement, la stratification est souvent réalisée à un point où une ou même moins d'une unité primaire est sélectionnée par strate.	Cela exige l'«effondrement» des strates semblables pour définir les nouvelles strates, telles que chacune contienne au moins deux sélections qui sont alors supposées comme indépendantes.
Les unités primaires sont parfois trop petites, variables ou bien encore inadaptées pour être utilisées directement dans la formule d'estimation de la variance.	Des unités calculables plus appropriées peuvent être définies par des techniques comme le groupement aléatoire des unités et la jonction ou l'union des unités à travers les strates.
Les échantillons sont normalement tirés sans remise.	Toutefois, pour des populations de grands échantillons à un faible taux, cela n'est presque jamais un problème dans les enquêtes sur la population.

Comment s'assurer que les erreurs d'échantillonnage peuvent être et sont calculées?

1. *Utilisez les échantillons probabilistes et mesurables.* L'échantillonnage probabiliste signifie l'utilisation des procédures aléatoires, dans le choix de l'échantillon, qui assigne une probabilité calculable non nulle à chaque élément de la population. La mesurabilité désigne un ensemble de critères pratiques qui permettent le calcul, à partir de l'échantillon même, des estimations ou des approximations valables de la variabilité de son échantillonnage.

2. *Définir convenablement la structure de l'échantillon pour assurer la «mesurabilité» de façon pratique.* Généralement, les procédures pratiques disponibles d'estimation de la variance nécessitent la présence de deux ou plusieurs sélections primaires dans chaque strate, tirée indépendamment et avec remise. Dans l'application pratique, il est parfois nécessaire de redéfinir la structure réelle pour adapter ce modèle. Les exemples ont été donnés dans le tableau ci-dessus. Une telle redéfinition des unités de calcul requiert nécessairement une expertise dans l'échantillonnage et de l'expérience.

3. *S'assurer que toutes les informations relatives à la structure de l'échantillon, nécessaires au calcul de la variance, sont à portée de main;* il est plus désirable de les coder comme partie intégrante du fichier de données de l'enquête.

4. *Utiliser des procédures générales et simples si possible, ce qui peut fournir des estimations de la variance utiles pour divers types de plans, de variables et de sous-classes.* L'utilisation de procédures plus spécifiques et plus complexes, même si elles sont plus précises, est secondaire à la production d'estimations au moins approximatives des erreurs d'échantillonnage de divers types d'estimation et de sous-groupes.

5. *Viser le calcul, l'analyse et la formalisation à grande échelle des erreurs d'échantillonnage.* Il est généralement insuffisant de limiter les calculs à quelques statistiques sélectionnées arbitrairement parmi le millier qui peut être produit pour différents variables, mesures et sous-groupes. Le modèle et l'ampleur des erreurs d'échantillonnage peuvent varier grandement pour différentes statistiques de la même enquête. Par conséquent, le calcul de routine et à grande échelle est généralement souhaitable. Toutefois, au même moment, il est nécessaire d'explorer les *modèles de variation des erreurs d'échantillonnage* (l'erreur type et les mesures dérivées comme le coefficient de variation, l'effet du plan de sondage et la corrélation intragroupe) pour diverses statistiques, calculer sur différents domaines d'échantillonnage et de sous-populations.

7.8. Mesures transportables de l'erreur d'échantillonnage

L'ampleur de l'erreur-type d'une statistique dépend de divers facteurs comme:

- la nature des estimations
- ses unités de mesures (échelle) et son ampleur

- variabilité des éléments dans la population (variance de la population)
- taille de l'échantillon
- la nature et la taille des unités d'échantillonnage
- structure de l'échantillon; procédures d'échantillonnage
- procédures d'estimation

Par conséquent, la valeur de l'erreur-type d'une statistique particulière est propre à la statistique concernée. Pour rapprocher l'erreur-type d'une statistique à celle d'une autre, il est nécessaire de décomposer l'erreur en composantes, dont l'effet de certains facteurs ci-dessus a été éliminé; c'est-à-dire que les composantes qui sont plus stables ou transportables d'un type de statistique ou du plan à une autre statistique ou plan. L'erreur-type d'une statistique telle qu'une moyenne est écrite dans les équations suivantes sous plusieurs formes, en termes de mesures transportables dans le sens ci-dessous.

Erreur type relative

$$\mathrm{se}(\bar{y}) = \bar{y}.\mathrm{rse}(\bar{y})$$

Ce terme désigne une erreur type d'une estimation, divisée par la valeur de l'estimation. Elle élimine l'effet sur l'erreur type de l'ampleur et l'échelle de mesure de l'estimation, mais elle dépend toujours d'autres facteurs comme la taille et le plan de l'échantillon.

L'erreur type dans un échantillon aléatoire simple (EAS) équivalent; variance de la population

L'erreur type d'une statistique estimée à partir d'un échantillon complexe peut être factorisée en deux parties:

(1) sr Erreur-type qu'on pourrait obtenir dans un échantillon aléatoire simple de même taille;

(2) deft Le facteur de fidélité, résumant l'effet des complexités du plan de sondage.

$$\mathrm{se}(\bar{y}) = \mathrm{deft}.\mathrm{sr}(\bar{y})$$

Le second élément (*sr*) est indépendant du plan de l'échantillon et a trait à la taille de l'échantillon d'une façon très simple:

$$\mathrm{sr}(\bar{y}) = s/\sqrt{n}$$

où l'écart-type *s* est une mesure de la variabilité de la population, indépendante du plan ou de la taille de l'échantillon. L'échelle de mesure peut aussi être supprimée en considérant le coefficient de variation, *cv*:

$$s = \bar{y}.\mathrm{cv}$$

L'écart-type est une mesure utile et fortement transportable. En outre, il peut être estimé d'une manière simple, indépendamment de la complexité du plan dans la plupart de situations pratiques. Par exemple, pour une proportion *p*

$$s^2 = \frac{n}{n-1}.p(1-p) \approx p(1-p)$$

alors que de façon plus générale, pour un ratio pondéré *r*, nous avons:

$$s^2 = \tfrac{n}{n-1}.\sum_i w_i z_i^2 / \sum_i w_i ; \quad z_i = (y_i - r.x_i)/\bar{x}$$

où

$$r = \sum_i w_i y_i / \sum_i w_i x_i ; \quad \bar{x} = \sum_i w_i x_i / \sum_i w_i$$

Le coefficient de variation est plus transportable, mais il n'est pas si utile quand le dénominateur dans sa définition est proche de zéro, comme c'est par exemple le cas pour les estimations des différences entre sous-classes. Par ailleurs, il n'y a aucun avantage à aller de *s* à *cv* dans le cas des proportions; en fait, le premier est préférable étant donné qu'il est symétrique (le même) pour la proportion (p) et son complément (1-p).

L'effet du plan de sondage et le taux d'homogénéité

L'*effet du plan de sondage*, *deft*^{2}, (ou sa racine carrée, *deft*, qui est parfois appelée *facteur de fidélité*) est une mesure sommaire compléte de l'effet sur l'erreur d'échantillonnage de diverses complexités du plan. En prenant le ratio de l'erreur-type de l'échantillon aléatoire réel (SRS), deft neutralise l'effet des facteurs courants des deux, telles que la taille de l'estimation et l'échelle de mesure, la variance de la population et la taille générale de l'échantillon. Toutefois, pour une variable donnée, son ampleur dépend toujours d'autres facteurs du plan.

Un principal facteur qui détermine la valeur *deft* est la taille de l'échantillon pris par UEP. Quand les tailles d'échantillon de l'UEP ne varient pas considérablement et que l'échantillon est essentiellement autopondérant, l'effet de ces tailles peut être isolée en considérant la mesure la plus transportable «*roh*»:

$$\text{deft}^2 = 1 + (\bar{b} - 1).\text{roh}$$

Dans la pratique, l'effet du plan de sondage d'une statistique est calculé en estimant sa variance (i) à partir du plan d'échantillonnage réel, et (ii) en supposant un simple échantillon aléatoire de même taille. Le ratio de ces deux quantités donne *deft*^{2}. Le paramètre *roh* peut être estimé à partir de ce *deft2* et du nombre moyen d'unités finales sélectionnées par UEP de l'échantillon, à l'aide de la formule ci-dessus.

Ce qui précède peut être affiné en isolant certaines sources de variation. Par exemple, en présence de tailles de l'échantillon de l'UEP variables, il est plus approprié de remplacer leur moyenne simple dans l'expression ci-dessus par la quantité

$$\overline{b'} = \overline{b}.(1 + cv^2) = \sum_i b_i^2 / \sum_i b_i$$

Une autre amélioration utile consiste à isoler l'effet sur le deft de D_w, de l'augmentation de la variance résultant des divergences arbitraires d'un plan autopondéré introduit plus tôt: (Paragraphe 7.1.3):

$$deft^2 = D_W^2.[1 + (\overline{b'} - 1).roh]$$

Il est plus important d'accumuler les informations sur les effets du plan. De telles informations sont essentielles pour améliorer le plan d'échantillonnage pour les enquêtes futures.

7.9. Quelle doit être l'ampleur de l'échantillon pris par grappe?

Il a été relevé au chapitre 2 que, à partir de l'examen des types d'échantillon en présence utilisés dans les enquêtes sur le travail des enfants (voir Tableaux 2.1 et 2.2), on trouve une marge étonnamment grande de variation dans les tailles des groupes (échantillon à prendre par zone) utilisé dans des enquêtes sur le travail des enfants. La marge de variation dépasse 10 à 50 ménages par zone. Si certaines de ces variations peuvent correspondre aux différentes circonstances nationales et aux différences relatives au type d'unités impliquées, il est probable que la plupart de ces variations ne soient pas basées sur les statistiques réelles ou les différences de coût.

Sur la base du calcul des erreurs d'échantillonnage et des effets des plans de sondage et les informations sur la répartition des coûts entre les zones et en leur sein, les plans d'échantillonnage des enquêtes sur le travail des enfants doivent être plus efficaces.

Un choix approprié du nombre d'unités finales (ménages, enfants) par zone d'échantillon dépend de la structure des coûts de l'enquête et du degré d'hétérogénéité (du variable d'intérêt au sein des groupes d'échantillon). Par structure du coût, nous entendons d'abord le coût relatif d'obtention des informations (notamment la collecte et le traitement des données) par unité finale, comparativement au coût moyen de dénombrement d'une zone d'échantillon (en plus du coût du dénombrement des unités finales). Le degré d'hétérogénéité dépend d'une gamme de facteurs, parmi lesquels le type d'unités utilisé comme unité d'échantillonnage primaire, la méthode de sous-échantillonnage à l'intérieur de ces UEP et – plus important encore – la variable d'enquête particulière qui est considérée.

L'expression simple suivante de la «taille de groupe optimale» est très utile pour indiquer les facteurs dont dépend le choix du nombre d'unités finales par zone d'échantillon et la manière de cette dépendance:

$$b_{OPT} = \left(\frac{C.(1 - roh)}{c. \cdot roh} \right)^{0.5}$$

Ici, C représente le coût par zone d'échantillon et c, le coût par unité finale de l'échantillon; roh, défini plus tôt, est la «corrélation intragroupe», mesurant le degré relatif d'hétérogénéité au sein des zones de l'échantillon. Pour les enquêtes dont les interviews coûtent (C), nous devrions avoir des tailles plus faibles par groupe. Lorsque le coût d'introduction de nouvelles zones dans l'échantillon est élevé (large C), l'échantillon optimum à prendre est augmenté (c'est-à-dire qu'il est préférable d'avoir quelques grappes plus grandes). Le paramètre roh est normalement un petit nombre (0,02-0,020) pour la plupart des variables, de telle sorte que le grand roh, l'échantillon optimum à prendre est plus faible. Remarquez que la racine carrée de ces paramètres est incluse dans l'expression ci-dessus. En d'autres termes, le résultat n'est heureusement pas trop sensible à la valeur des paramètres choisis. De bons compromis suffisent souvent dans la pratique.

Références

- Adames, Y.; Aquino Cornejo, M.; Castillo, R. et Rodriguez, A. 2003. *National report on the results of the child labour survey in Panama* (San Jose, OIT).
- Arnold-Talbert, E. et Constanza-Vega, L. 2003. *Child Labour in Belize: A Statistical Report 2001* (San Jose, OIT et Central Statistical Office, Government of Belize).
- Deming, W.E. 1960. *Sampling design in business research* (New York, John Wiley).
- Department of Census and Statistics of Sri Lanka. 1999. *Child Activity Survey Sri Lanka 1999* (Colombo, Department of Census and Statistics, Ministry of Finance and Planning).
- Dobles Trejos, C. et Pisoni, R. 2003. *National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica* (San Jose, OIT).
- Fox, K. 2004. *Report of Youth Activity Survey 2002* (Kingston, Statistical Institute of Jamaica).
- Ghana Statistical Service. 2003. *Ghana Child Labour Survey* (Accra, Ghana Statistical Service).
- IPEC. 2004. (Genève, OIT).
- IPEC. 2004. *National Report on the Results of the Child and Adolescent Labour Survey in Costa Rica, 2003* (San Jose, OIT).
- Kish, L. 1965. *Survey sampling* (New York, John Wiley).
- Kish, L. 1980. “Design and estimation for domains” dans *The Statistician,* vol. 29.
- Kish, L. 1987. *Small area statistics: An international symposium* (New York, John Wiley).
- Marschatz, A. 2004. *Report on the results of the National Child labour Survey in the Dominican Republic* (San Jose, OIT et State Department of Labour of the Dominican Republic).
- Ministry of Public Service, Labour and Social Welfare. 1999. *National Child Labour Survey, Country Report* (Harare, Ministry of Public Service, Labour and Social Welfare, Zimbabwe).
- National Institute of Statistics of Cambodia. 1997. *Report on Child Labour in Cambodia 1996* (Phnom Penh, National Institute of Statistics, Cambodge).
- National Statistical Office of Mongolia. 2004. *Report on National Child Labour Survey 2002-2003* (Ulaanbaatar, National Statistical Office of Mongolia).
- Orkin, F.M. 2000. *Survey of Activities of Young People* (SAYP) (Pretoria, Statistics South Africa).

Portugal Ministry of Labour and Solidarity (MTS). 1998. *Child labour in Portugal: Social Characterisation of School Age Children and Their Families* (Lisbonne, Portugal Ministry of Labour and Solidarity (MTS).

Purcell, N.J. et Kish, L. 1980. “Postcensal estimates for local areas (or domains)”. dans *International Statistical Review*, vol. 48.

Rao, J.N.K. 2003. *Small area estimation* (New York, John Wiley).

Shaik, K.R. 2000-2001. *Child Labour in Yemen* (Sana'a, Ministry of Labour of Yemen).

SIS. 1997. *Child Labour 1994* (Ankara, State Institute of Statistics Prime Ministry, Turquie, 1997).

State Statistics Committee of Ukraine. 2001. *Child Labour Ukraine 1999, Statistical Bulletin* (Kyiv, OIT et State Statistics Committee of Ukraine).

Thompson, S.K. 1990. “Adaptive cluster sampling” dans *Journal of the American Statistical Association.*

—. 1992. *Sampling* (New York, John Wiley).

Thompson, S.K. et Seber. G.A.F. 1996. *Adaptive Sampling* (New York, John Wiley).

Thompson, S.K. et Collins, L.M. 2002. “Adaptive sampling in research on risk-related behaviours” dans *Drug and Alcohol Dependence* vol. 68.

Programme international pour
l'abolition du travail des enfants (IPEC)

Organisation internationale du Travail
4, route des Morillons
CH-1211 Genève 22
Suisse

Tel.: +41 (0) 22 799 8181
Fax: +41 (0) 22 799 8771
Courriel: ilo@ilo.org
Site web: www.ilo.org/ipec

9 789222 225323