ST/ESA/STAT/SER.F/98

Département des affaires économiques et sociales
Division de statistique

Études méthodologiques Série F N° 98

Guide pratique pour la conception d'enquêtes sur les ménages

Nations Unies
New York, 2010

Département des affaires économiques et sociales

Le Département des affaires économiques et sociales du Secrétariat de l'Organisation des Nations Unies est un intermédiaire essentiel entre les politiques économiques, sociales et environnementales à l'échelle mondiale et l'action nationale. Ce département exerce ses activités dans trois domaines interdépendants : i) il recueille, établit et analyse un large éventail de données et d'informations économiques, sociales et environnementales dans lesquelles les États Membres de l'Organisation puisent pour examiner des problèmes courants et s'inspirer des possibilités d'action; ii) il facilite les négociations entre États Membres au sein de nombreux organes intergouvernementaux sur le choix d'actions communes à mener pour faire face à l'émergence de problèmes mondiaux; et iii) il conseille les gouvernements intéressés sur les moyens de passer des cadres d'action élaborés dans le contexte des conférences et des réunions au sommet tenues sous l'égide de l'Organisation des Nations Unies à des programmes à l'échelon national et, par le biais de l'assistance technique, il aide à renforcer les capacités nationales.

Note

Les appellations employées dans la présente publication et la présentation des données qui y figurent n'impliquent de la part du Secrétariat de l'Organisation des Nations Unies aucune prise de position quant au statut juridique des pays, territoires, villes ou zones ou de leurs autorités, ni quant au tracé de leurs frontières ou limites.

L'expression « pays ou zone » employée dans la présente publication s'entend aussi des territoires et des villes.

Les appellations « pays développé », « pays en développement » ou « pays les moins avancés » sont employées à des fins statistiques et n'expriment pas nécessairement une opinion quant au stade de développement de tel pays ou de telle zone.

Les cotes des documents de l'Organisation des Nations Unies se composent de lettres majuscules et de chiffres. La simple mention d'une cote dans un texte signifie qu'il s'agit d'un document de l'Organisation.

ST/ESA/STAT/SER.F/98

ISBN 978-92-1-261215-7

Publication des Nations Unies
Numéro de vente : F.06.XVII.13

Préface

L'objectif principal du *Guide pratique pour la conception d'enquêtes sur les ménages* est de réunir dans une seule et même publication les principales indications touchant la préparation d'enquêtes par sondage auxquelles puissent se référer les statisticiens, chercheurs et analystes nationaux appelés à réaliser des enquêtes sur les ménages dans leurs pays respectifs. Ce guide utilise des méthodes méthodologiquement rationnelles fondées sur la théorie statistique impliquant l'utilisation d'un échantillonnage probabiliste à chacune des phases du processus de sélection. Une enquête sur les ménages bien conçue et bien exécutée peut permettre de rassembler les informations nécessaires, d'une qualité et d'une exactitude suffisantes, rapidement et à relativement peu de frais.

Le contenu de cette publication peut également être utilisé comme guide de formation élémentaire à la conception d'enquêtes par sondage par les diverses institutions de formation à la statistique qui offrent des cours de statistiques appliquées, et en particulier des cours méthodologiques.

En outre, cette publication a été élaborée pour compléter les autres ouvrages consacrés à la méthodologie des enquêtes par sondage publiés par l'Organisation des Nations Unies, comme la récente publication intitulée *Enquêtes sur les ménages dans les pays en développement et les pays en transition*[1] et la série d'études publiées dans le cadre du Programme de mise en place de dispositifs nationaux d'enquêtes sur les ménages (NHSCP).

Plus spécifiquement, les objectifs de ce guide sont les suivants :

a) Présenter, en les regroupant, les concepts fondamentaux et les procédures et méthodes à appliquer pour concevoir des échantillons destinés principalement à la réalisation d'enquêtes nationales sur les ménages, l'accent étant mis sur les aspects appliqués de la conception d'échantillons de ménages;

b) Offrir des indications pratiques aux praticiens pour les aider à concevoir et à mener à bien des enquêtes;

c) Illustrer la corrélation entre la conception des échantillons et la collecte, l'estimation, le traitement et l'analyse des données;

d) Mettre en relief la nécessité d'identifier et de réduire, dans les enquêtes sur les ménages, les *erreurs dues à des facteurs autres que l'échantillonnage.*

S'il sera utile pour les usagers de ce guide d'être familiarisés avec les techniques d'échantillonnage, il pourra suffire d'avoir une connaissance générale des concepts statistiques et mathématiques pour pouvoir l'utiliser et l'appliquer sans guère d'assistance, voire aucune. En effet, l'un des principaux objectifs du guide est non pas de mettre l'accent sur les aspects théoriques de l'échantillonnage mais plutôt de présenter sous une forme pratique des indications qui puissent être immédiatement utilisées. En cas de besoin, cependant, ces indications sont accompagnées d'une explication de leur fondement théorique. Des rudiments d'algèbre devraient être suffisants pour suivre aisément les indications données et les appliquer. Aussi a-t-il été fourni de nombreux exemples pour illustrer les concepts, les méthodes et les techniques.

1 Études méthodologiques, n° 96 (publication des Nations Unies, numéro de vente : F.05.XVII.6).

Plusieurs experts ont contribué à l'élaboration du guide. M. Anthony Turner, consultant spécialisé dans l'échantillonnage, a rédigé les chapitres 3, 4 et 5 et a revu la version finale de l'ensemble du document; M. Ibrahim Yansaneh, chef adjoint de la Division du coût de la vie de la Commission de la fonction publique internationale, a rédigé les chapitres 6 et 7; et M. Maphion Jambwa, statisticien du Secrétariat de la Communauté de développement de l'Afrique australe, a rédigé le chapitre 9.

M. Jeremiah Banda, de la Division de statistique du Secrétariat de l'ONU, qui a fait fonction de rédacteur en chef et de coordonnateur technique du projet, est l'auteur des chapitres 1, 2 et 8 et de l'annexe I. Mme Clare Menozzi a aidé à éditer l'avant-projet de différents chapitres, et Mme Bizugenet Kassa a apporté un concours inappréciable au secrétariat du projet, et Mme Pansy Benjamin a assuré l'harmonisation des formats.

Les premiers chapitres ont été examinés par un groupe d'experts organisé par la Division de statistique qui s'est réuni à New York du 3 au 5 décembre 2003. L'on trouvera la liste des participants à cette réunion à l'annexe II. Le guide a également été soumis à un examen par les pairs de M. Alfredo Bustos, de Mme Ana Maria Landeros et de M. Eduardo Rios, de l'Institut national de la statistique, de la géographie et de l'informatique (INEGI) du Mexique, qui ont formulé de précieuses observations.

Paul Cheung
Directeur,
Division de statistique,
Département des affaires économiques et sociales
Organisation des Nations Unies

Table des matières

Chapitre 4
Cadres d'échantillonnage et échantillons-maîtres

Chapitre 5
Documentation et évaluation de la conception des échantillons

Chapitre 6
Construction et utilisation des pondérations d'échantillonnage

Chapitre 7
Estimation des erreurs d'échantillonnage dans les données d'enquête

FIGURES

Chapitre I

Sources des données utilisées pour les statistiques sociales et démographiques

1.1. Introduction

1. Les enquêtes sur les ménages sont au nombre des trois principales sources de statistiques sociales et démographiques dans de nombreux pays. Les recensements de la population et du logement constituent également d'importantes sources de statistiques sociales, mais ils sont habituellement réalisés à intervalles assez longs d'environ dix ans. Les dossiers administratifs constituent la troisième source des données. Pour la plupart des pays, cependant, cette source de données est plus développée pour les statistiques de la santé et les statistiques de l'état civil que pour les statistiques sociales. Les enquêtes sur les ménages peuvent avantageusement remplacer le recensement pour obtenir des données récentes et sont une formule plus utile et plus commode que les dossiers administratifs. Ces enquêtes sont utilisées pour rassembler des données sociodémographiques détaillées et variées concernant les conditions dans lesquelles vivent les populations, leur bien-être, les activités auxquelles elles se livrent et leurs caractéristiques démographiques ainsi que les éléments culturels qui influent sur les comportements et sur les transformations sociales et économiques. De telles enquêtes, cependant, n'interdisent pas d'avoir recours, à titre complémentaire, aux données provenant d'autres sources, comme les recensements et les dossiers administratifs.

1.2. Sources de données

2. Ces trois principales sources de données sociales et démographiques, si elles sont bien planifiées et bien exécutées, ou, dans le cas de dossiers administratifs, s'ils sont bien établis, peuvent se compléter pour former un programme intégré de collecte et de compilation de données. Les statistiques sociales et démographiques sont en effet essentielles à la planification et au suivi des programmes de développement socioéconomique. Des statistiques sur la composition de la population, par âge et par sexe, et sur sa répartition géographique sont au nombre des données les plus élémentaires qui sont nécessaires pour décrire une population et/ou un sous-groupe de population. Ces caractéristiques fondamentales définissent le contexte à l'intérieur duquel peuvent être étudiées d'autres informations importantes sur des phénomènes sociaux comme l'éducation, les handicaps, la participation à la population active, la santé, la situation nutritionnelle, la criminalité, la fécondité, la mortalité et les migrations.

1.2.1. Enquêtes sur les ménages

3. Les enquêtes sur les ménages sont devenues au cours des 60 à 70 dernières années l'une des principales sources de données permettant d'expliquer les phénomènes sociaux. Elles sont parmi les méthodes de collecte de données les plus souples. En théorie, presque tout sujet en rapport avec

la population peut être analysé par le biais d'enquêtes sur les ménages. Ainsi, utiliser les ménages comme unités d'échantillonnage secondaires est chose commune dans le contexte de la plupart des stratégies d'échantillonnage géographique (voir chapitres 3 et 4 du guide). Pour une enquête par sondage, l'on sélectionne une partie de la population, et des observations sont faites ou des données sont rassemblées à son sujet pour être ensuite extrapolées à l'ensemble de la population. Comme, dans une enquête par sondage, les enquêteurs ont moins de travail et plus de temps est réservé à la collecte des données, la plupart des questions peuvent être traitées plus en détail que dans le cadre d'un recensement. En outre, comme il n'est pas nécessaire d'avoir autant d'enquêteurs sur le terrain, il est possible de recruter des personnes mieux qualifiées et de leur dispenser une formation plus intensive que pour un recensement. La réalité est que les recensements ne permettent pas de rassembler toutes les données dont un pays a besoin; les enquêtes sur les ménages sont par conséquent un moyen de rassembler continuellement les informations supplémentaires ainsi que les nouvelles informations requises. Grâce à leur souplesse, les enquêtes sur les ménages constituent un excellent moyen de rassembler les données nécessaires à l'élaboration de statistiques qu'il serait autrement difficile de se procurer.

1.2.1.1. Types d'enquêtes sur les ménages

4. Beaucoup de pays ont institué des programmes d'enquêtes sur les ménages qui comprennent des enquêtes périodiques et des enquêtes ad hoc. À cet égard, il est bon que le programme d'enquêtes sur les ménages soit incorporé au système national intégré de collecte de données statistiques. Pendant les périodes intercensitaires, les enquêtes sur les ménages peuvent être l'élément de ce système à utiliser pour compiler des statistiques sociales et démographiques.

5. Le Programme de mise en place de dispositifs nationaux d'enquêtes sur les ménages (NHSCP) a été un vaste effort visant à aider les pays en développement à établir des dispositifs nationaux d'enquête et des moyens statistiques nécessaires pour rassembler les informations socioéconomiques et démographiques requises concernant le secteur des ménages. Ce programme a été réalisé pendant près de 14 ans, de 1979 à 1992. Lorsqu'il s'est achevé, une cinquantaine de pays y avaient participé à un moment ou à un autre. Son principal résultat a été la promotion et l'adoption par les pays de programmes d'enquêtes intégrées, continues et polyvalentes sur les ménages. En outre, le Programme a encouragé le renforcement des capacités en matière d'enquêtes par sondage, surtout dans les pays d'Afrique.

6. Plusieurs types d'enquêtes sur les ménages peuvent être menées pour rassembler des données sociales et démographiques : enquêtes spécialisées, enquêtes à phases multiples, enquêtes à champs multiples et enquêtes longitudinales. Le type d'enquête à choisir dépendra de différents facteurs, dont la question à étudier, les ressources disponibles et les considérations logistiques.

7. Les enquêtes spécialisées portent sur des questions ou des sujets ponctuels comme l'emploi du temps ou la situation nutritionnelle. Ces enquêtes peuvent être périodiques ou ad hoc.

8. Les enquêtes à phases multiples consistent à rassembler des informations statistiques en plusieurs étapes, chacune préparant la suivante. La phase initiale porte habituellement sur un échantillon plus nombreux que les suivantes. Son but est de filtrer les unités d'échantillonnage à la lumière de certaines caractéristiques de façon à déterminer si ces unités pourront être utilisées lors des phases ultérieures. Ces enquêtes constituent un moyen d'un bon rapport coût-efficacité de cerner la population cible qui servira à rassembler des informations détaillées sur une question spécifique lors des

phases ultérieures. De telles enquêtes sont utilisées notamment pour étudier des questions comme les handicaps ou la prévalence des enfants orphelins.

9. Les enquêtes à champs multiples ont pour but de rassembler des données sur plusieurs sujets simultanément. Cette approche est généralement plus économique que la réalisation de plusieurs enquêtes portant chacune sur une question distincte.

10. Les enquêtes longitudinales, quant à elles, ont pour but de rassembler des données concernant les mêmes unités d'échantillonnage tout au long d'une certaine période. Les intervalles peuvent être mensuels, trimestriels ou annuels. Le but de ces enquêtes est de mesurer l'évolution de certaines caractéristiques de la population considérée sur une période déterminée. Le principal problème que soulève ce type d'enquête est le taux élevé d'attrition des déclarants. L'effet de conditionnement peut également poser un problème.

1.2.1.2. Avantages et limitations des enquêtes sur les ménages et des recensements

11. Les enquêtes sur les ménages ne sont pas aussi onéreuses que les recensements, mais elles peuvent néanmoins être fort coûteuses si les résultats doivent être produits séparément pour des circonscriptions administratives relativement réduites comme une province ou un district. À la différence d'un recensement, dans le cas duquel des données seront rassemblées pour des millions de ménages, une enquête par sondage est habituellement circonscrite, pour des considérations de coût, à un échantillon de quelques milliers de ménages, ce qui limite sérieusement la possibilité pour de telles enquêtes de produire des données fiables pour des régions d'étendue réduite. La corrélation entre les dimensions de l'échantillon et la fiabilité des données pour des régions et des domaines réduits est étudiée dans les chapitres suivant.

12. Les enquêtes sur les ménages ont les avantages suivants sur les recensements :

a) Comme indiqué ci-dessus, le coût total d'un enquête est généralement inférieur à celui d'un recensement, ce dernier exigeant des ressources humaines, financières, logistiques et matérielles considérables. Un échantillon probabiliste, s'il est sélectionné et analysé comme il convient, permettra d'obtenir des résultats exacts et fiables qui pourront servir de base à des extrapolations à l'ensemble de la population. Pour certaines estimations comme le taux synthétique de fécondité, par conséquent, un recensement n'est pas absolument indispensable;

b) D'une manière générale, les enquêtes par sondage donnent des informations statistiques de meilleure qualité car, comme on l'a vu, elles permettent de recruter des enquêteurs mieux qualifiés et mieux formés. Le suivi peut également être meilleur car les superviseurs sont habituellement bien formés et le ratio entre les superviseurs et les enquêteurs peut parfois être de 1 à 4. En outre, l'on peut utiliser du matériel technique plus perfectionné pour prendre des mesures physiques lorsque celles-ci sont nécessaires. Dans le cas d'un recensement, la qualité des données se trouve parfois affectée par l'envergure même de l'opération, laquelle peut entraîner des défaillances de la qualité à différentes étapes et ainsi une incidence élevée d'erreurs autres que des erreurs d'échantillonnage;

c) Une enquête par sondage offre une gamme de choix plus large et plus de souplesse qu'un recensement pour ce qui est du degré de détail de l'étude et du nombre de points que peut viser le questionnaire. Il peut ne pas être possible de rassembler des informations de caractère plus spécialisé lors d'un recensement car le nombre de spécialistes ou la quantité de maté-

riels qui seraient nécessaires pour réaliser l'étude serait prohibitif. Il ne serait pas possible, par conséquent, de pondérer l'alimentation et des autres mesures dans une étude sur la nutrition. De même, il ne serait pas possible non plus de soumettre tous les membres d'une population à un examen médical pour déterminer, par exemple, l'incidence de l'infection par le VIH/sida. L'on peut en outre ajouter à une enquête par sondage des questions qu'il serait assez difficile de mesurer avec un recensement.

13. Les enquêtes par sondage se prêtent mieux à la collecte de données à l'échelle nationale et au niveau de domaines géographiques relativement vastes sur des questions qui doivent être analysées en profondeur comme les aspects multidimensionnels des handicaps, les dépenses des ménages, l'activité de la population active et la criminalité, à la différence des recensements, qui servent à rassembler des informations de caractère assez général portant sur des domaines restreints.

14. D'une manière générale, les enquêtes sur les ménages ont entre autres avantages le fait qu'il est utilisé des instruments de collecte de données suffisamment souples pour pouvoir poser un grand nombre de questions sur des sujets très divers et que sont estimés des paramètres comparables à ceux qui sont mesurés au moyen de recensements de la population et du logement.

1.2.2. Recensements de la population et du logement

15. Un recensement de la population, qui sera ci-après certainement dénommé un recensement, englobe tout le processus consistant à rassembler, compiler, évaluer et diffuser des données démographiques, sociales et autres couvrant, à un moment déterminé, tous les habitants d'un pays ou d'une ou plusieurs parties bien délimitées d'un pays. Un recensement constitue l'une des principales sources de statistiques sociales et a l'avantage évident de dégager des données fiables — c'est-à-dire des données qui ne sont pas affectées par des erreurs d'échantillonnage — pour des unités géographiques restreintes. Un recensement constitue le moyen idéal de rassembler des informations sur les effectifs, la composition et la répartition dans l'espace de la population ainsi que sur ses caractéristiques socioéconomiques et démographiques. D'une manière générale, un recensement a pour but de rassembler des informations sur chacun des membres d'un ménage et sur chaque série de localités de résidence, habituellement pour l'ensemble du pays ou pour des régions bien circonscrites du pays.

1.2.2.1. Principales caractéristiques d'un recensement classique de la population et du logement

16. Un recensement classique de la population et du logement présente les caractéristiques suivantes :

a) Chacun des membres de la population et chaque lieu de résidence sont dénombrés séparément et leurs caractéristiques sont enregistrées séparément;

b) L'objectif est de couvrir l'ensemble de la population d'un territoire clairement défini, c'est-à-dire de couvrir toutes les personnes qui s'y trouvent et/ou tous les résidents habituels, selon que le dénombrement est *de facto* ou *de jure.* En l'absence de registre de population et de dossiers administratifs complets, les recensements sont la seule source de statistiques étroitement localisées;

c) L'énumération est généralement aussi simultanée que possible dans l'ensemble du pays. Toutes les personnes et tous les logements sont énumérés pour la même période de référence;

d) Les recensements sont généralement menés à intervalles prédéterminés. La plupart des pays réalisent des recensements tous les dix ans, d'autres tous les cinq ans. Cela permet de disposer d'informations comparables à intervalles fixes.

1.2.2.2. Utilisations des résultats des recensements

17. En ce qui concerne les résultats des recensements :

a) Les recensements fournissent des informations sur les effectifs, la composition et la répartition dans l'espace de la population ainsi que sur ses caractéristiques démographiques et sociales;

b) Les recensements sont une source de statistiques étroitement localisées;

c) Les secteurs d'énumération utilisés pour les recensements sont la principale source de cadres d'échantillonnage pour les enquêtes sur les ménages. Les données rassemblées lors des recensements sont souvent utilisées comme informations auxiliaires aux fins de la stratification des échantillons ainsi que pour améliorer les estimations résultant des enquêtes sur les ménages.

1.2.2.3. Principales limitations des recensements

18. En raison de sa couverture géographique à nulle autre pareille, le recensement est habituellement une importante source de données de référence sur les caractéristiques de la population. Il n'est donc pas possible d'étudier en détail un grand nombre de questions. Il se peut que le recensement ne soit pas la meilleure source d'informations détaillées concernant, par exemple, l'activité économique, de telles informations exigeant des questions et des investigations détaillées.

19. Comme le recensement fait très largement appel à des déclarants indirects, il ne permet pas toujours de rassembler des informations sur des caractéristiques qui peuvent n'être connues que de l'intéressé, comme l'occupation, les heures travaillées, le revenu, etc.

20. Des recensements de la population ont été réalisés dans beaucoup de pays au cours des quelques dernières dizaines d'années. Pendant la série de 2000 (1995-2004), par exemple, 184 pays et territoires ont procédé à des recensements.

1.2.3. Dossiers administratifs

21. Des statistiques sociales très diverses sont compilées à partir des dossiers administratifs, qui sont en quelque sorte le sous-produit de l'activité de l'administration. L'on peut en citer comme exemples les statistiques sur la santé tirées des registres des hôpitaux, les statistiques concernant l'emploi tirées des services de placement, les statistiques provenant des registres de l'état civil, ou les statistiques concernant l'éducation tirées des registres sur la fréquentation scolaire tenus par les Ministères de l'éducation. La fiabilité des statistiques tirées des dossiers administratifs dépend de leur complétude ainsi que de la cohérence des définitions et des concepts.

22. Si les dossiers administratifs permettent d'obtenir des données à très peu de frais, il est rare que de tels systèmes soient suffisamment établis dans les pays en développement, ce qui signifie que, dans la majorité des cas, ces données sont inexactes. Même si les processus d'enregistrement sont continus aux fins de l'administration, la compilation de statistiques est le plus souvent un souci secondaire pour la plupart des organisations, de sorte que la qualité des données en souffre. En outre, il n'est

habituellement pas tenu compte des normes statistiques qui doivent être respectées en ce qui concerne par exemple la normalisation des concepts et des définitions ou la complétude de la couverture.

23. Dans la plupart des pays, les informations provenant des dossiers administratifs n'ont souvent qu'un contenu limité dans la mesure où elles sont utilisées surtout à des fins juridiques ou administratives. Les systèmes de registres de l'état civil sont des exemples de systèmes administratifs mis en place dans de nombreux pays. Cependant, les pays n'ont pas tous réussi dans cet effort. Les pays qui disposent de systèmes complets de registres de l'état civil peuvent produire périodiquement des rapports sur des questions comme le nombre de naissances vivantes par sexe, par date et par lieu de naissance, le nombre de décès par âge, par sexe et par lieu et cause du décès, les mariages et les divorces, etc.

24. Un registre de population tient une base de données sur la vie de chaque habitant et de chaque ménage du pays. Le registre est continuellement mis à jour lorsque les caractéristiques d'un individu et/ou d'un ménage changent. Ces registres, s'ils sont combinés à d'autres registres sociaux, peuvent être une riche source d'information. Les pays qui ont mis au point de tels systèmes sont notamment l'Allemagne, le Danemark, la Norvège, les Pays-Bas et la Suède. Pour la plupart de ces pays, les recensements sont fondés sur le système de registres.

25. Dans de nombreux pays en développement, les dossiers administratifs tenus dans le cadre de différents programmes sociaux peuvent être une source de données utiles et économiques et une proposition attrayante, mais ils ne sont guère développés. Les dossiers administratifs n'ont souvent qu'un contenu limité et, habituellement, n'ont pas la même souplesse que les enquêtes sur les ménages pour ce qui est des concepts ou du degré de détail possible. Lorsque tel est le cas, il est très difficile de les utiliser en même temps que d'autres sources étant donné que les concepts ne sont pas normalisés et que la couverture des systèmes de classification est sélective ou incomplète.

1.2.4. Complémentarité des trois sources de données

26. L'on a vu comment les recensements, les enquêtes et les systèmes de dossiers administratifs peuvent être utilisés de concert. La présente section examine plus particulièrement la possibilité de combiner les informations provenant de différentes sources de façon complémentaire. Si cela est intéressant, c'est parce qu'il importe de réduire les coûts des recensements et des enquêtes, de faciliter les réponses, d'obtenir des données à un niveau inférieur de désagrégation et de maximiser l'utilisation des données disponibles dans le pays.

27. Comme les recensements ne peuvent pas être répétés fréquemment, les enquêtes sur les ménages permettent de mettre à jour certaines des informations provenant des recensements, surtout au plan national ou au niveau d'autres grands domaines. Dans la plupart des cas, un recensement ne porte que sur des sujets relativement simples, et le nombre de questions est habituellement limité. Les informations provenant des recensements peuvent par conséquent être complétées par des informations détaillées sur des sujets complexes provenant des enquêtes sur les ménages, qui présentent l'avantage de porter sur des échantillons plus petits et d'être souples.

28. Fréquemment, les recensements et les enquêtes sur les ménages sont complémentaires. Rassembler pendant le recensement des informations sur d'autres questions parmi un échantillon de ménages est un moyen à la fois efficace et économique d'élargir la portée du recensement pour collecter les statistiques sociales qui sont de plus en plus nécessaires. L'utilisation de méthodes et de techniques

d'échantillonnage permet de produire des données dont on a un besoin urgent avec une précision acceptable lorsqu'il serait difficile, pour des raisons de temps et de coût, d'obtenir ces données au moyen d'une énumération complète.

29. Le recensement fournit également le cadre d'échantillonnage, l'infrastructure et les moyens statistiques et les statistiques de référence qui sont nécessaires pour réaliser des enquêtes sur les ménages. Il n'est pas inhabituel d'identifier un échantillon de ménages dans le contexte d'un recensement pour rassembler des informations sur des sujets plus complexes comme les handicaps, la mortalité maternelle, l'activité économique et la fécondité.

30. Les recensements facilitent les enquêtes sur les ménages en offrant des cadres d'échantillonnage : le recensement donne une liste explicite de toutes les unités, comme secteurs d'énumération, communément utilisées comme unités primaires dans le processus de sélection. De plus, certaines informations auxiliaires provenant d'un recensement peuvent être utilisées pour concevoir au mieux une enquête sur les ménages. En outre, ces informations auxiliaires peuvent servir à améliorer les estimations des échantillons au moyen d'analyses de régression et d'estimations des ratios et ainsi d'améliorer la précision des estimations.

31. Pour intégrer les sources de données, il importe d'identifier clairement les unités d'énumération et d'adopter des unités géographiques cohérentes pour la collecte de statistiques provenant de différentes sources. En outre, il est essentiel d'adopter des définitions, des classifications et des concepts communs pour les différentes sources de données, y compris les dossiers administratifs.

32. Les données provenant des enquêtes sur les ménages peuvent être utilisées aussi pour contrôler la couverture et le contenu des recensements afin de déterminer ainsi l'ordre de grandeur et la direction des erreurs. Les enquêtes post-énumération ont été utilisées à cette fin lors de la série de recensements de 2000 en Zambie et au Cambodge pour évaluer les erreurs de couverture. Les données provenant des recensements peuvent, de même, être utilisées pour évaluer certains résultats des enquêtes.

33. Les estimations hautement localisées, qui sont de plus en plus demandées, constituent un domaine dans lequel les données provenant des enquêtes et des dossiers administratifs sont utilisées pour obtenir simultanément des estimations. Les estimations localisées directes classiques ne sont pas suffisamment précises car, lorsque le secteur géographique est restreint, les dimensions des échantillons sont rarement suffisantes. Les estimations localisées sont fondées sur toute une série de méthodes statistiques utilisées pour produire des estimations lorsque les estimations traditionnelles provenant des enquêtes ne sont pas fiables ou ne peuvent pas être calculées. L'on utilise notamment des modèles qui fournissent un lien avec des secteurs localisés au moyen de données supplémentaires ou auxiliaires, comme les dernières données provenant des recensements de la population. Ces méthodes reposent donc essentiellement sur l'idée consistant à emprunter et à combiner les avantages relatifs des différentes sources de données afin de produire des estimations plus exactes et plus fiables.

34. Dans les pays où les registres de l'état civil sont bien développés, l'on peut utilement exploiter les données provenant des recensements et des enquêtes en même temps que les données provenant des dossiers administratifs. Lors du recensement de la population réalisé en 1990 à Singapour, par exemple, les enquêteurs disposaient préalablement d'informations de base provenant des dossiers administratifs pour chaque membre du ménage. Cette approche a permis d'accélérer les entrevues et de réduire les coûts de l'énumération. Comme un recensement fondé sur les registres de l'état civil ne permet d'obtenir de données que sur les effectifs totaux de la population et sur ses principales

caractéristiques, des informations sur ses caractéristiques socioéconomiques détaillées sont rassemblées par sondage.

35. Les données provenant des dossiers administratifs peuvent être utilisées pour vérifier et évaluer les résultats des enquêtes et des recensements. Par exemple, dans les pays où les systèmes de registres de l'état civil sont complets, les données relatives à la fécondité et à la mortalité tirées des recensements peuvent être comparées à celles qui proviennent du système de registres de l'état civil.

1.2.5. Conclusions

36. Les enquêtes sur les ménages, les recensements et les dossiers administratifs doivent être considérés comme complémentaires. Cela signifie que, dans toute la mesure possible, il faudra utiliser des définitions et des concepts communs pour la planification des recensements et des enquêtes. Il conviendra également de vérifier périodiquement les dossiers administratifs pour veiller à ce que les définitions et les concepts employés soient uniformes.

37. Le programme d'enquêtes sur les ménages doit faire partie d'un système national intégré de collecte de données statistiques, y compris au moyen de recensements et de dossiers administratifs de façon à pouvoir réunir les statistiques sociodémographiques nécessaires.

Références et autres lectures

Ambler, R. *et al.* (2001). Combining unemployment benefits data and LFS to estimate ILO unemployment for small areas: an application of the modified Fay-Herriot method. Document établi en vue de la réunion de l'Institut international de statistique, Séoul.

Banda, J. (2003). Current status of social statistics: an overview of issues and concerns. Document présenté à la réunion du Groupe d'experts consacrée à la détermination du champ d'application des statistiques sociales organisée par la Division de statistique du Secrétariat de l'ONU en collaboration avec le Groupe de Sienne sur les statistiques sociales. New York, 6-9 mai 2003.

Bee-Geok, L. et K. Eng-Chuan (2001). ESA/STAT/AC.88/05, 7 avril. Combining survey and administrative data for Singapore's census of population 2000. Document établi en vue de la réunion de l'Institut international de statistique, Séoul.

Kiregyera, B. (1999). *Sample Surveys: With Special Reference to Africa.* Kampala, PHIDAM Enterprises.

Organisation des Nations Unies (1982). *Programme de mise en place de dispositifs nationaux d'enquêtes sur les ménages*. Non-sampling errors in household surveys. Sources, Assessment and Control. Version préliminaire. DP/UN/INT-81-041/2. New York, Département de la coopération technique pour le développement et Bureau de statistique du Secrétariat de l'ONU.

_____ (1984). *Handbook of Household Surveys* (édition révisée). Études méthodologiques, n° 31, numéro de vente : E.83.XVII.13.

_____ (1998). *Principles and Recommendations for Population and Housing Censuses,* Révision 1. Études statistiques, n° 67/Rev.1, numéro de vente : E.98.XVII.8.

_____ (2001). *Principles and Recommendations for a Vital Statistics System,* Révision 2. Numéro de vente : E.01.XVII.10. ST/ESA/STAT/SER.M/19/Rev.2.

_____ (2002). Technical report on collection of economic characteristics in population censuses. New York et Genève, Division de statistique, Département des affaires économiques et sociales, et Bureau de statistique, Bureau international du Travail. ST/ESA/STAT/119. En anglais seulement.

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology* (Statistique Canada, Ottawa), vol. 25, n° 2, p. 175-186.

Singh, R. et N. Mangat (1996). *Elements of Survey Sampling.* Boston, Massachusetts, Kluwer Academic Publishers.

Statistique Canada (2003). *Survey Methods and Practices.* Ottawa.

Whitfold, D. et J. Banda (2001). Post enumeration surveys (PES's): are they worth it? Document présenté au Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid–decade Assessment and Future Prospects. New York, 7-10 août 2001, organisé par la Division de statistique du Département des affaires économiques et sociales du Secrétariat de l'ONU. Symposium 2001/10. En anglais seulement.

Chapitre 2

Planification et exécution des enquêtes

1. Bien que l'accent soit mis, dans le présent guide, sur l'aspect échantillonnage des enquêtes sur les ménages, il importe de donner un aperçu général de la planification et de la réalisation d'une enquête sur les ménages afin de replacer les chapitres et sections concernant l'échantillonnage dans le contexte approprié. Il existe d'innombrables manuels et guides qui traitent très en détail de la planification et de la réalisation d'enquêtes sur les ménages, et le lecteur est instamment engagé à s'y référer pour plusieurs informations. Nombre des principaux points, cependant, sont mis en relief et décrits brièvement dans le présent chapitre, y compris les principales caractéristiques de la planification et de la réalisation des enquêtes, sauf pour ce qui est de la conception et de la sélection des échantillons, question qui est abordée aux chapitres 3 et 4 et à l'annexe I.

2.1. Planification des enquêtes

2. Il faut, si l'on veut qu'une enquête donne les résultats souhaités, accorder une attention particulière aux préparatifs qui doivent précéder le travail sur le terrain. À ce propos, toutes les enquêtes exigent une préparation soigneuse et judicieuse si l'on veut qu'elles soient couronnées de succès. Cependant, l'étendue du travail de planification variera selon le type d'enquête et les informations requises. Comme il faut disposer de ressources et d'un temps suffisants pour élaborer un plan adéquat (voir la figure 2.1), il n'est pas inhabituel que la planification d'une enquête complexe prenne jusqu'à deux ans (pour une discussion détaillée de la planification des enquêtes, voir Organisation des Nations Unies, 1984).

La figure 2.1 illustre le calendrier qui peut être suivi.

2.1.1. Objectifs d'une enquête

3. Il importe au plus haut point de définir clairement les objectifs d'une enquête dès le début du projet. Il doit être établi une description statistique clairement formulée des informations recherchées, de la population et de la couverture géographique. Il faut également, à ce stade, indiquer comment le résultat doit être utilisé. Le statisticien chargé de la préparation de l'enquête devra définir les objectifs à la lumière du budget alloué. Tenir dûment compte des contraintes budgétaires, en effet, facilitera la planification et la réalisation de l'enquête.

4. Dans certains cas, les objectifs de l'enquête ne sont pas définis explicitement. Par exemple, un service d'enquêtes pourra être invité à réaliser une étude de l'activité dans le secteur non structuré. Si l'objectif de l'enquête n'est pas défini clairement, il appartiendra au statisticien et au responsable de l'enquête de définir le secteur non structuré en termes opérationnels aux fins de l'enquête et de cerner

Figure 2.1
Calendrier des activités liées à une enquête sur les ménages dans le pays X

Tâche	Durée	2005 Fév.	Mars	Avril	Mai	Juin	Juil.	Août	Sept.	Oct.	Nov.	Déc.	2006 Janv.	Fév.	Mars	Avril	Mai	Juin	Juil.	Août	Sept.	Oct.	Nov.	Déc.
1. Réunion avec les parties prenantes	1 mois																							
2. Activités préparatoires	3 mois																							
3. Conception initiale du questionnaire	2 mois																							
4. Envoi du questionnaire aux membres des comités d'usagers	1 mois																							
5. Incorporer les suggestions convenues au questionnaire	1 mois																							
6. Rédiger les manuels de l'enquêteur et du superviseur	2 mois																							
7. Imprimer les questionnaires et les manuels	1 mois																							
8. Planifier le test préliminaire	1 mois																							
9. Former les enquêteurs et les superviseurs	2 mois																							
10. Réaliser le test préliminaire	1 mois																							
11. Réviser le questionnaire (si besoin est)	1 mois																							
12. Réviser le questionnaire (si besoin est)	1 mois																							
13. Concevoir l'échantillon	1 mois																							
14. Concevoir et tester le(s) programme(s) d'entrée des données	1 mois																							
15. Créer et tester le(s) programme(s) de nettoyage des données	1 mois																							
16. Collecte des données	3 mois																							
17. Saisie des données	4 mois																							
18. Concevoir et tester le plan de tabulation	2 mois																							
19. Nettoyage des données (vérifications de la portée et de la cohérence)	2 mois																							
20. Affectation des chapitres aux auteurs	2 mois																							
21. Traitement des données conformément au plan	1 mois																							
22. Calcul des erreurs d'échantillonnage, entre autres estimations	1 mois																							
23. Distribution de la tabulation aux auteurs	1 mois																							
24. Analyse et rédaction du rapport	2 mois																							
25. Compilation et tabulation du rapport final	2 mois																							

en détail les activités économiques spécifiques qui répondent de plus près aux besoins de l'institution ayant demandé la réalisation de l'enquête. Il y a lieu de mentionner à ce propos qu'une enquête dont les objectifs sont restés ambigus et vagues risque fort de souffrir d'une proportion élevée d'erreurs autres que d'échantillonnage.

5. Il importe au plus haut point que les parties prenantes, c'est-à-dire les divers usagers et producteurs de statistiques, soient associées à la définition des objectifs de l'enquête ainsi que de sa portée et de sa couverture. Ces consultations aident en effet à dégager des consensus ou à parvenir à des compromis au sujet des données qui sont requises, de la forme sous laquelle elles doivent être présentées, des niveaux de désagrégation, des stratégies de diffusion et de la fréquence des opérations de collecte de données.

6. Certaines des enquêtes réalisées comportaient des objectifs très précis. Par exemple, l'enquête pilote sur la population active réalisée en Zambie en 1983 avait les objectifs ci-après :

a) Rassembler des informations sur les effectifs et la composition de la population travaillant dans le secteur structuré;

b) Évaluer la demande et l'offre de main-d'œuvre;

c) Servir de base à des projections de la main-d'œuvre pour des groupes professionnels déterminés;

d) Aider à planifier l'expansion de l'éducation dans des domaines revêtant une importance capitale pour le développement économique.

7. Il y a lieu de noter qu'il faut commencer par définir clairement les objectifs pour pouvoir identifier les questions auxquelles il faudra trouver des réponses statistiques.

2.1.2. Contexte de l'enquête

8. Pour planifier une enquête, il faut définir les secteurs géographiques et la population cible. Pour une enquête sur les revenus et les dépenses des ménages, par exemple, l'enquête pourra porter sur les régions urbaines et peut-être exclure les régions rurales.

9. Pour définir le contexte de l'enquête, il faudra identifier avec précision la population pour laquelle il faut sélectionner un échantillon. Dans le cas d'une enquête sur le revenu et les dépenses des ménages, le contexte des unités primaires serait les secteurs d'énumération (il s'agirait des unités géographiques visées réparties, par exemple, sur l'ensemble du pays) et les unités secondaires seraient les ménages vivant dans les secteurs d'énumération sélectionnés (la question des grappes est discutée plus en détail aux chapitres 3 et 4; voir également l'annexe I pour la définition des grappes).

10. Il y a lieu de souligner toutefois que, dans la pratique, la population cible est un peu plus réduite que celle qui constitue le contexte, c'est-à-dire l'univers. Habituellement, l'on restreint la population cible, pour différentes raisons. Dans le cas de certaines enquêtes, certains ménages de militaires vivant dans les casernes pourraient être exclus de l'enquête. Dans le cas des enquêtes sur la population active, les enfants n'ayant pas atteint un âge spécifié peuvent apparaître comme faisant partie des ménages interrogés, mais non de la population active.

11. Il importe de noter que lorsque la population réelle diffère de la population cible, les résultats ne s'appliqueront qu'à la population spécifique dont il a été pris un échantillon. Comme on le verra

au chapitre 4, il importe de construire, pour chaque étape du processus de sélection, des cadres complets et s'excluant mutuellement.

2.1.3. Informations à rassembler

12. L'on peut établir, parmi la liste des questions appelant une réponse statistique, une liste de points pouvant fournir des informations factuelles sur les thèmes à l'étude. Il importe de ne jamais perdre de vue que certaines des données nécessaires peuvent être tirées de sources existantes. Pour établir la liste, il faut envisager d'y inclure des questions supplémentaires liées aux thèmes principaux. Dans le cadre d'une enquête sur l'emploi et les salaires, par exemple, il peut être réuni des informations supplémentaires sur l'âge, le sexe et le niveau d'instruction des déclarants. Ces informations permettraient d'élucider des questions connexes et, ainsi, enrichiraient l'analyse.

13. Il convient d'ajouter qu'il y aura lieu, lors de la planification de l'enquête, de préparer un plan de tabulation. Les tableaux en blanc devront être distribués pour observations et amélioration éventuelle.

2.1.4. Budget de l'enquête

14. Le budget de l'enquête identifie les ressources requises. Il est nécessaire pour appuyer et guider la réalisation de l'enquête et l'établissement du calendrier suivant lequel les résultats devront être présentés. Les estimations de coûts doivent être aussi détaillées que possible. Il faut donc bien comprendre toutes les activités détaillées que suppose la réalisation de l'enquête. Le budget fait apparaître les dépenses de personnel, les dépenses d'équipement et tous les autres postes de dépenses. S'il y a un plafond de ressources à ne pas dépasser, ce qui est habituellement le cas, le budget global doit s'inscrire à l'intérieur du cadre prédéterminé. Il est bon aussi, lors de l'établissement du budget, de suivre les indications générales données par l'institution qui financera l'enquête afin de faciliter ainsi l'approbation du projet de budget. S'il faut s'écarter du budget prescrit, les autorisations voulues devront être demandées aux organisations intéressées. Les demandes de ressources doivent être établies dès que possible. D'une manière générale, le budget dépendra pour une large part de la conception de l'enquête, du degré de précision requis et de la couverture géographique. L'on trouvera à la figure 2.2 ci-après un modèle pouvant être utilisé pour calculer les coûts.

15. Il importe pour l'organisation chargée de réaliser l'enquête de mettre en place un système efficace de maîtrise des coûts. Il y a dans la plupart des enquêtes de grande envergure un sérieux risque de perte de contrôle sur le décaissement des fonds dès que commencent les activités sur le terrain. En pareilles circonstances, des ressources considérables tendent à être dépensées pour des activités sans rapport avec les activités principales. Un contrôle judicieux des coûts aide à suivre les dépenses effectives à la lumière des dépenses prévues et du travail effectivement réalisé. Il importe au plus haut point que le personnel de direction responsable de l'enquête fasse en sorte qu'il soit dûment rendu compte de l'utilisation qui est faite des ressources. Cela rehausse considérablement la crédibilité de l'organisation chargée de l'enquête.

Figure 2.2
Exemple d'un modèle de calcul des coûts d'un programme d'enquête sur les ménages

Activité	Estimation des unités de travail (mois-personne, sauf indication contraire)	Coût unitaire (unité monétaire par mois-personne, sauf indication contraire	Coût estimatif total (en unités monétaires)
I. PLANIFICATION ET ACTIVITÉS PRÉPARATOIRES			
A. Planification initiale et suivi ultérieur (personnel de direction)			
B. Sélection et spécification du sujet			
1. Planification du sujet			
2. Préparation des plans de tabulation			
3. Services de secrétariat et autres services			
C. Conception de l'enquête			
1. **Conception initiale : structure de l'enquête, population cible, procédures d'échantillonnage, méthodes de collecte de données, etc.** (personnel professionnel)			
2. Élaboration des matériels d'échantillonnage :			
a) Matériels cartographiques (à supposer que de tels matériels soient disponibles pour les recensements) :			
Dépenses de personnel			
Cartes et fournitures			
b) Établissement de listes des ménages (2 000 zones d'énumération):			
Dépenses de personnel (principalement enquêteurs)			
Frais de voyage			
c) Sélection et préparation des échantillons sur la base des listes de terrain			
D. Conception et impression des questionnaires et autres documents			
1. Personnel professionnel			
2. Services de secrétariat et autres services			
3. Frais d'impression (après essais préliminaires)			
E. Essai préliminaire			
1. Planification : personnel professionnel :			
a) Préparatifs initiaux			
b) Analyse des résultats et révision des matériels			
2. Superviseurs de terrain :			
a) Dépenses de personnel			
b) Frais de voyage			
3. Enquêteurs :			
a) Dépenses de personnel			
b) Frais de voyage			

Activité	Estimation des unités de travail (mois-personne, sauf indication contraire)	Coût unitaire (unité monétaire par mois-personne, sauf indication contraire	Coût estimatif total (en unités monétaires)
F. Préparation d'instructions et de matériels de formation pour utilisation sur le terrain			
1. Personnel professionnel			
2. Services de secrétariat et autres services			
3. Frais de reproduction			
G. Activités de planification diverses (par exemple relations publiques et publicité			
H. Totaux partiels, par composante			
1. Personnel de direction			
2. Personnel professionnel			
3. Personnel technique			
4. Personnel de service			
5. Frais de voyage			
6. Impression			
7. Cartographie et divers			
TOTAL PARTIEL			
II. OPÉRATIONS SUR LE TERRAIN			
A. Formation des superviseurs de terrain			
1. Dépenses de personnel			
2. Hébergement et repas			
3. Frais de voyage			
B. Formation des enquêteurs			
1. Superviseur			
2. Enquêteurs :			
a) Dépenses de personnel			
b) Frais de voyage			
C. Collecte de données (y compris contrôle de la qualité)			
1. Superviseur :			
a) Dépenses de personnel			
b) Frais de voyage			
2. Enquêteurs :			
D. Administration sur le terrain			
1. Direction des activités sur le terrain			
2. Frais de voyage			
3. Dépenses diverses (par exemple contrôle et expédition des matériels)			
E. Totaux partiels, par composante			
1. Personnel professionnel			

Activité	Estimation des unités de travail (mois-personne, sauf indication contraire)	Coût unitaire (unité monétaire par mois-personne, sauf indication contraire	Coût estimatif total (en unités monétaires)
2. Personnel technique			
3. Personnel de service			
4. Frais de voyage			
5. Indemnité de subsistance			
6. Entrevues			
7. Divers			
TOTAL PARTIEL			
III. TRAITEMENT DES DONNÉES			
A. Planification des systèmes			
B. Programmation des systèmes informatiques			
C. Codage			
1. Codage initial			
2. Contrôle de la qualité			
3. Supervision			
D. Opérations de transfert sur disque			
1. Saisie initiale			
2. Contrôle de la qualité			
2. Supervision			
E. Services informatiques (y compris opérateurs et dépenses de maintenance			
F. Frais divers de traitement (fournitures, etc.)			
G. Totaux partiels, par composante			
1. Personnel professionnel			
2. Personnel technique			
3. Personnel de contrôle de la qualité			
4. Personnel de service			
5. Informatique			
6. Divers			
TOTAL PARTIEL			
IV. RÉVISION ET PUBLICATION DES DONNÉES			
A. Personnel professionnel			
B. Frais de publication			
V. DIRECTION ET COORDINATION DE L'ENQUÊTE (supervision continue de toutes les activités)			
VI. TOTAL PARTIEL			
VII. ÉTUDES D'ÉVALUATION ET RECHERCHES MÉTHODOLOGIQUES (ce chiffre peut être estimé comme étant de 10% du total global)			

Activité	Estimation des unités de travail (mois-personne, sauf indication contraire)	Coût unitaire (unité monétaire par mois-personne, sauf indication contraire	Coût estimatif total (en unités monétaires)
VIII. FRAIS GÉNÉRAUX (ce chiffre peut être estimé comme étant de 15% du total global pour les frais d'administration, la location des locaux, les fournitures, etc.)			
IX. TOTAL			

Source : Nations Unies (1984).

2.2. Exécution des enquêtes

2.2.1. Méthodes de collecte des données

16. Plusieurs méthodes peuvent être utilisées pour rassembler des données, dont une observation et une mesure directes, l'envoi d'un questionnaire par le courrier et des entrevues par téléphone et en personne.

17. *Observation et mesure directes :* L'observation et la mesure directes constituent la méthode idéale, car celle-ci est habituellement plus objective. Il n'y a pas à se préoccuper de trous de mémoire et d'un risque de subjectivité de la part des déclarants ou des enquêteurs. L'observation directe a été utilisée par exemple dans des domaines comme les suivants :

a) Certains aspects des enquêtes sur la consommation alimentaire;

b) Des enquêtes sur les prix, les enquêteurs pouvant alors acheter le produit et en enregistrer le prix.

18. Cette méthode, bien qu'utile, a pour inconvénient d'exiger un investissement considérable de temps et d'argent. Le plus souvent, les enquêteurs ont besoin d'un certain matériel. L'expérience a montré que la méthode de l'observation et de la mesure directes est la plus utile et la plus pratique lorsque les dimensions de l'échantillon ou des populations sont relativement réduites.

19. *Envoi d'un questionnaire par le courrier :* L'envoi d'un questionnaire par le courrier est une méthode assez économique et rapide. Au stade de la collecte des données, la principale dépense à prévoir est l'affranchissement. Une fois que le questionnaire a été conçu et imprimé, il est envoyé par la poste aux déclarants (c'est-à-dire aux personnes qui sont censées remplir le questionnaire). L'on tient pour acquis que les déclarants savent lire et écrire, étant censés remplir eux-mêmes le questionnaire. Cependant, tel peut ne pas être le cas, surtout dans les pays en développement, où les taux d'alphabétisation demeurent peu élevés. Le principal inconvénient de cette méthode est le taux élevé de non-réponse (c'est-à-dire une proportion élevée de déclarants qui ne remplissent pas le questionnaire et/ou ne répondent qu'à certaines questions), ce qui peut être dû à la complexité du questionnaire utilisé. Cependant, une absence de réponse peut également être due à l'apathie. Dans certains cas, le taux de réponse au questionnaire est élevé, mais beaucoup de questions restent sans réponse.

20. Pour essayer d'améliorer le taux de réponse, il peut s'avérer nécessaire d'envoyer des rappels, mais il est bon de sélectionner un sous-échantillon de non-déclarants et de les interroger personnellement. Cela peut être nécessaire lorsque les caractéristiques des unités non déclarantes sont tout

à fait différentes de celles des déclarants (voir la question de la stratification et de ses avantages au chapitre 3 et à l'annexe I). En l'occurrence, les unités déclarantes et non déclarantes sont considérées comme deux domaines qui doivent être pondérés différemment lors de la préparation des estimations (la pondération est une question examinée plus en détail dans les chapitres suivants, et surtout au chapitre 6). Pour améliorer le taux de réponse, le questionnaire envoyé doit être attrayant, bref et aussi simple que possible. Inclure une enveloppe préaffranchie pour le renvoi du questionnaire peut aussi contribuer à améliorer le taux de réponse.

21 Pour que cette méthode puisse donner des résultats satisfaisants, il faut également disposer d'un cadre d'échantillonnage qui soit aussi à jour que possible, en particulier pour ce qui est de l'adresse des déclarants. L'organisation chargée de l'enquête doit également avoir l'assurance que les déclarants sont capables de remplir eux-mêmes le questionnaire.

22 En résumé, certains des avantages et des limitations des enquêtes menées par l'envoi d'un questionnaire sont les suivants :

Avantages :

a) Elles sont moins chères;

b) L'échantillon peut être très disséminé;

c) Le parti pris de l'enquêteur se trouve éliminé;

c) Elles sont rapides.

Limitations :

a) Le taux de non-réponse est habituellement élevé;

b) Les réponses aux questions doivent être prises pour argent comptant, car il n'est pas possible de demander des éclaircissements;

c) Dans une enquête sur les attitudes, il est difficile de déterminer si le déclarant a répondu aux questions sans l'aide de quelqu'un d'autre;

d) La méthode est utile seulement lorsque les questionnaires sont assez simples et ne se prêtent donc pas à des enquêtes complexes.

23 *Entrevue personnelle :* C'est la méthode la plus communément utilisée dans les pays en développement pour rassembler des données au moyen d'enquêtes de grande envergure. Outre que le taux de réponse est généralement élevé dans le cas d'entrevues personnelles, cette méthode est appropriée car, dans certains de ces pays, les taux d'analphabétisme sont élevés. Selon cette formule, les enquêteurs s'adressent à des déclarants sélectionnés pour rassembler des informations ou pour poser des questions. Le principal avantage de cette méthode est que les enquêteurs peuvent expliquer aux déclarants pourquoi il convient qu'ils répondent aux questions et peuvent expliquer les objectifs de l'enquête. En outre, il est plus facile de rassembler des informations statistiques sur des questions conceptuellement complexes qui pourraient susciter des réponses ambiguës si un questionnaire était envoyé par la poste.

24 La méthode de l'entrevue personnelle présente néanmoins certaines limitations, par exemple :

a) Des enquêteurs différents peuvent interpréter différemment les questions et introduire ainsi un biais dans les résultats de l'enquête, très rares étant ceux qui se réfèrent toujours au manuel d'instructions;

b) Certains enquêteurs peuvent, lorsqu'ils demandent des éclaircissements, suggérer des réponses aux déclarants;

c) Les caractéristiques personnelles des enquêteurs, par exemple leur âge, leur sexe et parfois même leur race, peuvent influencer les attitudes des déclarants;

d) Il se peut que les enquêteurs lisent mal les questions, devant s'occuper simultanément de poser des questions et d'enregistrer les réponses.

25. Collectivement, les limitations susmentionnées sont les principales sources de ce que l'on appelle la distorsion liée à l'enquêteur, laquelle, comme l'ont montré les études, peut entraîner de sérieuses erreurs autres que d'échantillonnage dans les enquêtes (voir la discussion concernant les erreurs autres que d'échantillonnage au chapitre 8).

26. Lorsque l'on interroge les déclarants, il faut tenir compte des points ci-après :

a) L'enquêteur doit bien comprendre l'objet de chaque question, tel qu'il est expliqué dans le manuel. Il importe que les enquêteurs se réfèrent constamment au manuel;

b) L'expérience a montré qu'il est préférable pour l'enquêteur de suivre l'ordre des questions figurant dans le questionnaire. Le plus souvent, il a été soigneusement réfléchi à l'ordre des questions en ayant à l'esprit la motivation des déclarants, les liens entre les sujets, la nécessité de rafraîchir la mémoire du déclarant et les questions les plus délicates;

c) Les enquêteurs doivent absolument s'abstenir de suggérer des réponses aux déclarants;

d) Toutes les questions doivent être posées. L'on peut ainsi réduire au minimum le nombre de questions auxquelles il n'est pas répondu. En outre, aucun point du questionnaire ne doit être laissé en blanc à moins qu'il ne s'agisse d'une question à sauter. Si une question est sans objet pour un déclarant déterminé, il conviendra d'en indiquer les raisons. Cette approche permet en effet au responsable de l'enquête d'avoir l'assurance que toutes les questions reflétées dans le questionnaire ont été administrées.

2.2.2. Conception du questionnaire

27. Une fois que les objectifs de l'enquête et le plan de tabulation ont été déterminés, l'on peut préparer le questionnaire. Celui-ci joue un rôle capital dans le processus d'enquête étant donné que c'est lui qui facilite le transfert de l'information de celui qui l'a (les déclarants) à ceux qui en ont besoin (les usagers). C'est l'instrument par l'entremise duquel les informations dont les usagers ont besoin sont exprimées en termes opérationnels et c'est la principale source des informations à entrer dans le système de traitement des données.

28 La longueur et la présentation du questionnaire sont deux questions qui doivent être étudiées très attentivement. Il est bon de concevoir le questionnaire au moment de la planification de l'enquête. Si le questionnaire doit être envoyé par la poste aux déclarants, il doit être attrayant et simple, ce qui pourra accroître le taux de réponse. D'un autre côté, un questionnaire devant être utilisé pour

que les enquêteurs puissent y porter les réponses données sur le terrain doit être suffisamment résistant pour survivre à sa manipulation.

29. Idéalement, le questionnaire devrait être conçu de manière à faciliter la collecte de données pertinentes et exactes. Pour améliorer l'exactitude des informations recueillies lors de l'enquête, il faudra s'attacher tout particulièrement à ordonner comme il convient les questions posées et à les rédiger avec soin. Le déclarant doit être motivé. Le questionnaire doit être suffisamment aéré pour que l'enquêteur ou le déclarant puisse lire facilement les questions. L'on ne saurait trop insister sur le fait que tout questionnaire doit être accompagné d'instructions claires.

30. L'équipe chargée de l'enquête devra veiller par conséquent à définir avec précision les données à rassembler et à bien spécifier comment les données nécessaires et les concepts connexes doivent être reflétés dans les questions qui seront posées. À ce propos, il faut habituellement faire un essai préliminaire pour mettre le questionnaire à l'épreuve, à moins que celui-ci ait été pleinement validé lors d'enquêtes précédentes.

31. En bref, un bon questionnaire doit :

a) Permettre de rassembler les informations exactes répondant au moment opportun aux besoins des usagers potentiels des données;

b) Faciliter le travail de collecte, de traitement et de tabulation des données;

c) Permettre de rassembler les données à peu de frais, c'est-à-dire éviter de collecter des informations non essentielles;

d) Permettre une analyse détaillée et utile et une utilisation productive des informations rassemblées.

32. Cela signifie que les questionnaires doivent être conçus de manière à produire des informations de la plus haute qualité possible, l'accent étant mis en particulier sur leur pertinence, leur exactitude et leur actualité. Pour que le processus puisse être mené à bien efficacement, il faut réduire au minimum le coût et la charge de travail que représente la collecte des informations requises.

2.2.2.1. Construction des questions

33. Les questions généralement posées dans les questionnaires d'enquête sont des questions ouvertes ou des questions fermées. Dans le cas d'une réponse ouverte, le déclarant donne sa propre réponse. Dans une enquête sur les attitudes, les déclarants peuvent être invités à définir ce qu'ils entendent par une bonne qualité de vie. Manifestement, les divers déclarants définiront comme ils l'entendent ce qui constitue une bonne qualité de vie. D'un autre côté, une question fermée oblige le déclarant à sélectionner une des réponses figurant sur la liste donnée par l'équipe d'enquêteurs. Voici quelques exemples de questions fermées :

Souffrez-vous d'une incapacité mentale permanente qui limite vos activités quotidiennes ?

❑ Oui ❑ Non

Comment évaluez-vous votre vision (même avec des lunettes ou des lentilles de contact, si vous en utilisez) ?

1. ❑ Absence de vision
2. ❑ Sérieuse difficulté permanente

3. ❑ Quelque difficulté permanente
4. ❑ Pas de difficulté.

34. Les questions fermées présentent l'avantage qu'elles : *a*) produisent des réponses plus uniformes; et *b*) peuvent être traitées facilement. Leur principale limitation est que les réponses possibles doivent être préparées par le concepteur de l'enquête, ce qui signifie que des réponses qui peuvent être importantes risquent de se trouver négligées. Dans la plupart des enquêtes, il est préférable de poser des questions ouvertes au sujet de sujets complexes et à propos des attitudes et des idées.

2.2.2.2. Libellé des questions

35 Les questions doivent être claires, précises et dépourvues d'ambiguïté. Il ne faut pas demander au déclarant de deviner ce que l'enquêteur cherche à extraire de lui. Les définitions et les concepts employés peuvent paraître évidents pour les enquêteurs, mais pas pour le déclarant. En conséquence, le déclarant pourra faire appel à son jugement pour répondre aux questions, de sorte que le résultat risque d'être une multitude d'erreurs autres que d'échantillonnage. Prenons un exemple simple. Dans de nombreux pays d'Afrique, surtout en milieu urbain, la question « Où habitez-vous ? » risque de susciter une confusion si le verbe « habiter » n'est pas clairement défini, car de nombreux déclarants interprètent cette question comme se référant au village dont ils sont initialement venus.

2.2.2.3. Questions « tendancieuses »

36. Les questions dites tendancieuses conduisent le déclarant à répondre aux questions d'une certaine façon. Autrement dit, la question tend à privilégier une certaine réponse. Exemple de question tendancieuse dans le contexte d'une enquête sur la santé: « Combien de fois par semaine buvez-vous plus de deux bouteilles de bière ? » Cette question force le déclarant à admettre qu'il boit de la bière, en fait, pas moins de deux bouteilles par jour. De telles questions tendent à fausser les réponses. Il importe d'éviter de créer des données : l'objectif est simplement de les rassembler.

2.2.2.4. Pertinence des questions

37. Un questionnaire doit servir à obtenir des informations qui seront utilisées pour étudier la situation dont il s'agit. Il importe donc au plus haut point que l'organisation chargée de l'enquête pose des questions pertinentes afin de brosser un tableau exact de la situation à l'étude. Les questions figurant dans le questionnaire doivent être pertinentes pour la plupart des déclarants. Par exemple, dans l'environnement rural qui caractérise la plupart des pays d'Afrique à l'heure actuelle, administrer un questionnaire encombré de questions sur les résultats obtenus au niveau de l'enseignement supérieur (universitaire) n'aurait aucun sens. De même, il n'y a pas lieu, dans une enquête sur la fécondité, d'inclure les femmes de moins de 10 ans et de leur poser des questions sur les enfants qu'elles ont eus ou sur le point de savoir si elles sont mariées, divorcées ou veuves, questions qui sont pertinentes pour des femmes ayant atteint un certain âge, mais pas pour des filles qui n'ont pas atteint l'âge de procréer.

2.2.2.5. Séquence des questions

38 Les thèmes évoqués dans le questionnaire doivent être rangés dans un ordre de nature à rafraîchir la mémoire du déclarant et à faciliter l'obtention d'informations exactes. Le mieux est que

les premières questions soient faciles, intéressantes et neutres afin de mettre le déclarant à l'aise et de l'encourager à poursuivre une entrevue, à laquelle, le plus souvent, il participe volontairement. Aujourd'hui, il est assez habituel aussi d'ordonner les questions posées lors d'enquêtes sur les ménages de manière à commencer par les questions visant à identifier l'unité d'échantillonnage, comme l'adresse, en posant ensuite les questions tendant à définir les caractéristiques du ménage et les personnes qui le composent, par exemple leurs caractéristiques démographiques. Enfin, l'on pose des questions détaillées qui constituent le but principal de l'enquête[1]. D'une manière générale, les questions délicates doivent être parmi les dernières à être posées. L'important, à ce stade, est que les questions, surtout celles qui appellent des réponses dépendant de la précédente, soient rangées dans un ordre logique.

2.2.3. Plan de tabulation et d'analyse

39. Il existe une technique utile qui aide le concepteur de l'enquête à mieux préciser les outils visant à rassembler les informations dont l'usager a besoin, telles qu'elles sont reflétées dans les questions posées ou dans les objectifs de l'enquête. Cette technique consiste à produire des plans de tabulation et des tableaux en blanc, qui sont des tableaux comprenant toutes les rubriques sauf les informations proprement dites. À tout le moins, le canevas de tabulation doit spécifier les titres des tableaux et des colonnes et identifier les variables techniques à indiquer, les variables de référence à utiliser aux fins de la classification et les groupes de population (objets, éléments ou unités visés par l'enquête) auxquels s'appliquent les divers tableaux. Il est bon aussi de faire apparaître de manière aussi détaillée que possible les différentes catégories de classifications, bien que celles-ci puissent être modifiées ultérieurement lorsque la répartition de l'échantillon sur les différentes catégories de réponse sera mieux connue.

40. L'importance d'un plan de tabulation apparaît clairement de différents points de vue. Par exemple, les projets de tableaux qui avaient été établis permettront de déterminer si les données à rassembler déboucheront sur des tableaux utiles en faisant apparaître non seulement ce qui manque, mais aussi ce qui est superflu. En outre, l'investissement supplémentaire de temps qui est fait dans l'élaboration de projets de tableaux est habituellement plus que compensé par le temps gagné, lors de la tabulation, sur la conception et la traduction des tableaux finalement utilisés.

41. Il existe également une étroite corrélation, à ne pas perdre de vue, entre le plan de tabulation et la conception de l'échantillon utilisé pour l'enquête. Par exemple, il n'est possible de décomposer les tableaux par région géographique que si l'échantillon est conçu de manière à permettre une telle ventilation. De même, du fait des dimensions de l'échantillon, il peut être nécessaire de limiter le nombre de cases dans les tableaux pour éviter que ceux-ci soient trop fragmentés. Parfois, le plan devra être modifié pendant le travail de tabulation. Il peut également être nécessaire de combiner plusieurs catégories pour réduire le nombre de cases vides, ou bien les constatations intéressantes découlant des données préliminaires pourront conduire à élaborer de nouveaux tableaux. La façon dont les données rassemblées lors de l'enquête sur les ménages seront utilisées pour répondre aux questions (atteindre les objectifs de l'enquête) peut être appelée, en termes plus généraux, « plan d'analyse des données ». Ce plan indique en détail quelles sont les données nécessaires pour réaliser les objectifs de l'enquête. Les concepteurs devront par conséquent s'y référer continuellement lorsqu'ils élabore-

[1] Voir Organisation des Nations Unies, 1984.

ront les détails du questionnaire. Il va peut-être sans dire que le plan d'analyse doit également être le principal point auquel il faut se référer pour analyser les résultats de l'enquête.

2.2.4. Exécution du travail sur le terrain

42. Dans la plupart des pays en développement, le travail sur le terrain est souvent très sérieusement entravé par le manque de ressources. Cependant, s'il faut mener une enquête, le travail sur le terrain doit être organisé et réalisé comme il convient pour que les ressources limitées qui sont disponibles puissent être utilisées au mieux. Si l'on veut que l'enquête soit couronnée de succès, les concepteurs devront bien comprendre les aspects conceptuels de son objet. En outre, les enquêteurs devront être particulièrement familiarisés avec les procédures qui, dans la pratique, leur permettront de rassembler des données exactes. Ainsi, il est toujours nécessaire de disposer sur le terrain d'une structure bien organisée et efficace.

2.2.4.1. Équipement et matériel

43. Dans beaucoup de pays en développement, il faut mobiliser à l'avance du matériel comme véhicules, bateaux, bicyclettes, etc., en état de marche. Il faut également constituer une réserve de pièces de rechange. S'ils disposent de véhicules et de bicyclettes, les chefs d'équipe, superviseurs et enquêteurs pourront se déplacer plus facilement et plus rapidement.

44. Des fournitures adéquates comme classeurs, crayons, taille-crayons, bloc-notes et carburant (pour les véhicules) devront être disponibles en quantité suffisante pendant l'enquête.

2.2.4.2. Gestion des opérations d'enquête

45. Une enquête de grande envergure est habituellement une opération exigeante et complexe. L'on ne saurait donc trop insister sur la nécessité pour toutes les activités d'être gérées judicieusement et de manière efficace et efficiente à tous les niveaux.

46. Il importe également de définir clairement les structures hiérarchiques entre le responsable de l'enquête et les enquêteurs. Il y a lieu de noter à ce propos que des formulaires de suivi de l'avancement de l'enquête se sont avérés utiles.

2.2.4.3. Publicité

47. Certaines enquêtes n'ont donné que des résultats limités par suite, entre autres, du taux élevé de non-réponse dû aux refus d'y participer. Les organisateurs doivent par conséquent prévoir certaines campagnes de publicité. L'expérience a montré en effet que la publicité peut beaucoup encourager la coopération des déclarants, même si certaines organisations de financement considèrent les dépenses de publicité comme un gaspillage de ressources.

48. L'on peut adopter différentes approches de la publicité, selon les circonstances. Dans les régions urbaines de certains pays, les publicités à la radio, à la télévision et dans les journaux peuvent compléter des affiches, tandis qu'il peut être préférable, en milieu rural, d'avoir recours à des messages à la radio ainsi qu'à des affiches.

49. Il peut être nécessaire aussi d'organiser des réunions avec les personnalités les plus influentes des régions sélectionnées afin de les informer des objectifs de l'enquête. En outre, ces personnalités

devraient être invitées à convaincre les habitants de leurs localités de répondre comme il convient aux questions des enquêteurs.

50. Avant d'entreprendre le travail sur le terrain, il importe de rendre publique en l'expliquant la décision qui se trouve à la base de l'enquête. Cette annonce devra, entre autres informations, indiquer les objectifs et la durée de l'enquête et les questions qui seront abordées.

2.2.4.4. Sélection des enquêteurs

51. L'enquêteur se trouve à l'interface avec les déclarants. Comme c'est le représentant de l'organisateur de l'enquête qui se trouve constamment en contact avec le déclarant, la façon dont il s'acquitte de son travail influe directement sur le succès de l'enquête. La sélection des enquêteurs doit par conséquent beaucoup retenir l'attention et être faite sérieusement. L'enquêteur doit être capable de communiquer efficacement avec le déclarant et posséder toutes les qualités nécessaires pour encourager le déclarant à fournir dans un délai raisonnable des informations exactes sur les points évoqués.

52. L'enquêteur devra avoir un niveau d'instruction correspondant au type de l'enquête à mener. En outre, il devra pouvoir enregistrer honnêtement les informations reçues, sans les « embellir ». Les enquêteurs sélectionnés devront suivre les instructions qui leur sont données et utiliser les définitions et les concepts reflétés dans le manuel.

53. Les procédures ci-après pourront aider à sélectionner les enquêteurs appropriés :

a) Les candidats devront remplir une demande indiquant notamment quels sont leur âge, leur situation conjugale, leur adresse, leur niveau d'instruction et leurs antécédents professionnels;

b) Les candidats présélectionnés pourront être appelés à subir un test d'intelligence et un autre test de calcul simple;

c) Comme, indépendamment des tests écrits, il faut habituellement avoir une entrevue avec les candidats, les entrevues devront être menées par un jury qui les note de manière indépendante. Certaines des qualifications à prendre en considération à cette fin seront notamment l'entregent, l'intérêt pour le travail, l'ouverture et la perspicacité.

54. Le travail sur le terrain peut être ardu et le déplacement pourrait être difficile, de sorte que les enquêteurs sélectionnés devront être prêts à travailler dans des conditions souvent rudes.

2.2.4.5. Formation des enquêteurs

55. Les enquêteurs sélectionnés devront recevoir une formation approfondie avant d'être affectés sur le terrain, essentiellement pour uniformiser les procédures d'enquête, ce qui est également nécessaire pour éviter que les enquêteurs n'interprètent différemment les définitions, les concepts et les objectifs de l'enquête et pour minimiser ainsi les distorsions dues aux enquêteurs.

56. La formation devra être dispensée par des instructeurs qualifiés qui devront manifestement être familiarisés avec les buts et les objectifs de l'enquête. De préférence, ils devront faire partie de l'équipe chargée de mener l'enquête.

57. Les enquêteurs devront recevoir des instructions très précises concernant les buts de l'enquête et la façon dont les résultats en seront utilisés. Pour être informés comme il convient des objectifs de

l'enquête, les enquêteurs devront être suffisamment familiarisés avec les concepts et les définitions utilisés dans le questionnaire.

58. Dans le cadre du processus de formation, les enquêteurs, en présence de l'instructeur, devront chacun, tour à tour, expliquer à leurs collègues les divers points du questionnaire. Des travaux pratiques devront être réalisés aussi bien en classe que sur le terrain. Par exemple, l'on pourra inviter les enquêteurs, tour à tour, à se poser entre eux des questions en classe, et l'on pourra ensuite les emmener sur le terrain, dans un quartier voisin, où les futurs enquêteurs pourront s'entretenir avec quelques ménages. L'instructeur devra toujours être présent pour guider les enquêteurs et les corriger. Après les entrevues sur le terrain, les futurs enquêteurs devront en discuter les résultats sous la direction de l'instructeur. Le programme de formation pourra amener le responsable de l'enquête à décider des enquêteurs qui devront recevoir une formation supplémentaire et de ceux dont les attitudes ne correspondent absolument pas à celles qu'exige l'emploi.

2.2.4.6. Supervision sur le terrain

59. Chacun s'accorde généralement à reconnaître qu'une formation est indispensable à un travail efficace et réussi sur le terrain. Cependant, une formation peut, en l'absence de supervision appropriée, ne pas donner les résultats souhaités. Le travail sur le terrain ne peut aboutir que si des cadres plus expérimentés et mieux qualifiés que les enquêteurs supervisent continuellement et efficacement leur travail. Les superviseurs devront être formés à tous les aspects de l'enquête. L'on ne saurait trop insister sur le fait que le superviseur est un lien important entre l'organisation de collecte de données et l'enquêteur. Il est censé organiser le travail des enquêteurs en arrêtant les affectations; il doit vérifier le travail accompli et veiller à ce que les enquêteurs se consacrent entièrement à l'enquête. En outre, si possible, il y a lieu de prévoir un ratio relativement élevé entre le personnel d'encadrement et les enquêteurs. Pour la plupart des enquêtes sur les ménages, l'idéal paraît être un superviseur pour quatre ou cinq enquêteurs, mais ce chiffre a simplement valeur d'indication.

2.2.4.7. Entrevues avec les personnes n'ayant pas répondu

60. Il y a inévitablement, dans toutes les enquêtes, des cas de non-réponse (voir le chapitre 8 pour les erreurs autres que d'échantillonnage). Il se peut que certains déclarants refusent de coopérer avec les enquêteurs ou que certaines questions ne soient pas posées. Lorsqu'une unité n'ayant pas répondu a été signalée au superviseur, celui-ci doit se mettre en rapport avec elle et essayer, grâce à sa plus grande expérience, d'obtenir les informations demandées. Comme le but de toute enquête est d'obtenir le taux de réponse le plus élevé possible, il est recommandé de rassembler cette information auprès d'un sous-échantillon du groupe initial de non-déclarants. L'effort est ainsi réorienté vers ce sous-échantillon, les questions devant alors, de préférence, être posées par des superviseurs.

2.2.4.8. Réduction du taux de non-réponse

61. Il importe, lors de la conception et de l'exécution d'une enquête sur les ménages, d'élaborer des méthodes de nature à maximiser le taux de réponse. Il faut bien comprendre l'importance qu'il y a à élaborer des procédures afin de réduire le nombre de refus, par exemple en offrant de revenir pour mener l'entrevue au moment qui convient le mieux au déclarant. En outre, les objectifs de l'enquête et l'utilisation qui en sera faite devront être soigneusement expliqués aux déclarants pour les encourager

à coopérer. Une garantie de confidentialité peut également aider à atténuer la crainte des déclarants que leurs réponses soient utilisées autrement qu'aux fins assignées à l'enquête.

62. Lorsque personne n'est pas à la maison, il conviendra de répéter les visites à différentes heures de la journée, jusqu'à quatre fois si nécessaire.

63. Il importe également d'éviter les problèmes qui se posent lorsque les unités d'échantillonnage sélectionnées ne peuvent pas être localisées, ce qui peut être une source importante de non-réponse. Le mieux, en présence de ce problème, est d'utiliser la dernière version disponible du cadre d'échantillonnage (voir le chapitre 4 pour un examen détaillé de cette question).

Références et autres lectures

Kiregyera, B. (1999). *Sample Surveys: With Special Reference to Africa.* Kampala, PHIDAM Enterprises.

Organisation des Nations Unies (1982). Programme de mise en place de dispositifs nationaux d'enquêtes sur les ménages. Non-sampling errors in household surveys: Sources, Assessment and Control. Version préliminaire. DP/UN/INT-81-041/2. New York, Département de la coopération technique pour le développement et Bureau de statistique du Secrétariat de l'ONU.

_____ (1984). *Handbook of Household Surveys* (édition révisée). Études méthodologiques, n° 31, numéro de vente : E.83.XVII.13.

_____ (1998). *Principles and Recommendations for Population and Housing Censuses,* Révision 1. Études statistiques, n° 67/Rev.1, numéro de vente : E.98.XVII.8.

_____ (2002). Technical report on collection of economic characteristics in population censuses. New York et Genève, Division de statistique, Département des affaires économiques et sociales, et Bureau de statistique, Bureau international du Travail. ST/ESA/STAT/119. (En anglais seulement.)

Statistique Canada (2003). *Survey Methods and Practices.* Ottawa.

Zanutto, E. et A. Zaslavsky (2002). Using administrative records to improve small area estimation: an example from the U.S. Decennial Census. *Journal of Official Statistic* (Statistics Sweden), vol. 18, n° 4, p. 559-576.

Chapitre 3

Stratégies d'échantillonnage

3.1. Introduction

1. Le chapitre 2, relatif à la planification des enquêtes, a donné un aperçu général des différentes phases des opérations d'enquête, mais le présent chapitre est le premier d'une série de chapitres consacrés exclusivement aux divers aspects de l'échantillonnage, ce qui est le thème principal du guide. Ce chapitre discute brièvement de l'échantillonnage probabiliste par opposition à l'échantillonnage non probabiliste et explique pourquoi c'est toujours le premier qu'il faut utiliser pour une enquête sur les ménages. L'on s'est étendu longtemps sur les dimensions de l'échantillon, c'est-à-dire sur les nombreux paramètres qui le déterminent et sur la façon de le calculer. Il est également suggéré un certain nombre de techniques pour garantir l'efficacité du processus d'échantillonnage, dont la stratification, l'échantillonnage en grappes et l'échantillonnage par phases, l'accent étant mis en particulier sur la conception d'échantillons en deux phases (voir les définitions et les descriptions de ces concepts au tableau 3.1 et à l'annexe I). Différentes formules sont suggérées et deux principaux types de conception qui ont été utilisés dans de nombreux pays sont décrits en détail. Le chapitre traite également des questions particulières que sont : *a*) l'échantillonnage en deux phases pour parvenir jusqu'aux populations « rares »; et *b*) la variation ou la tendance de variation du ratio entre l'échantillon et les estimations. Le chapitre s'achève sur quelques brèves recommandations.

3.1.1. Aperçu général

2. Presque tous les échantillons utilisés pour les enquêtes sur les ménages, aussi bien dans les pays en développement que dans les pays développés, sont difficiles à concevoir en raison de leurs caractéristiques : multiples phases, strates et grappes. En outre, le fait que les enquêtes nationales par sondage sur les ménages ont souvent une portée générale et portent sur de multiples sujets présentant de l'intérêt pour l'État aggrave encore leur complexité. Le présent guide fait par conséquent une place particulière aux stratégies d'échantillonnage multiphases.

3. Pour parvenir au résultat souhaité, un échantillon bien conçu, qui est un arrangement symphonique, doit combiner harmonieusement de nombreux éléments. L'échantillon doit être sélectionné en plusieurs *phases* de sorte qu'il soit possible de déterminer avec précision où les entrevues seront menées et de choisir efficacement les ménages à interroger. La conception doit être *stratifiée* de manière à ce que l'échantillon effectivement sélectionné soit réparti comme il convient sur plusieurs sous-régions géographiques et sous-groupes de population. Le plan d'échantillonnage doit faire appel à des grappes, c'est-à-dire à des unités habituellement définies sur une base géographique parmi lesquelles les ménages sont sélectionnés, afin de maintenir les coûts dans des limites gérables. Simultanément, il y a lieu d'éviter une *mise en grappes* excessive car cela peut affecter la fiabilité (voir la section 3.3.5 pour l'effet de la mise en grappes). Les *dimensions* de l'échantillon doivent tenir compte des exigences concurrentes à respecter, de façon à concilier coût et précision. Les dimensions de l'échantillon doivent également être telles que les usagers puissent obtenir les informations dont ils ont un besoin urgent concernant différents domaines, à savoir des sous-groupes de population ou des sous-régions

géographiques. La conception de l'échantillon doit tendre à garantir le maximum d'exactitude à deux égards importants : premièrement, le *cadre d'échantillonnage* utilisé (ou élaboré) doit être aussi complet, correct et à jour que possible et, deuxièmement, les techniques de sélection des échantillons doivent être utilisées de manière à minimiser les distorsions involontaires parfois causées au niveau de la réalisation de l'enquête. Un système d'auto-évaluation doit également être incorporé à la conception de l'échantillon, lequel doit permettre d'estimer les *erreurs d'échantillonnage* de façon à aider les usagers à évaluer la fiabilité des principaux résultats. Les erreurs d'échantillonnage peuvent découler d'une estimation des caractéristiques de la population cible fondée sur des données concernant une partie seulement de la population plutôt que son intégralité.

4. Une enquête a essentiellement pour objectif de permettre, sur la base d'un échantillon aléatoire, d'extrapoler concernant la population cible. Ce faisant, l'enquêteur/chercheur essaie habituellement d'estimer telle ou telle caractéristique inconnue de la population. Parmi les caractéristiques ou paramètres les plus communément utilisés, l'on peut citer les totaux, les moyennes, les proportions et les variances. Par exemple, si $Y_1, Y_2, Y_3, \ldots \ldots Y_N$ sont les valeurs d'une variable de la population,

la moyenne de la population est $\bar{Y} = \frac{1}{N}\sum Y_i$ **(3.1)**

la variance de la population est $\sigma^2 = \frac{1}{N}\left(\sum Y^2 - N\bar{Y}^2\right)$ **(3.2)**

Le plus souvent, des estimations de l'échantillon sont utilisées pour estimer les paramètres de la population. Par exemple, la moyenne et la variance de l'échantillon pour un échantillon de dimensions *n* pour un échantillon aléatoire simple sélectionné au moyen d'une procédure de remplacement sont données par les équations :

$$\bar{Y} = \frac{1}{N}\sum Y_i \quad \textbf{(3.3)}$$

$$s^2 = \frac{1}{n-1}\left(\sum y_i^{\,2} - n\bar{y}^2\right) \quad \textbf{(3.4)}$$

où $y_1, y_2, y_3, \ldots \ldots y_n$ sont les valeurs de la variable *y* pour *n* unités de l'échantillon. Dans une enquête par sondage, le chercheur calcule les variances des variables aléatoires sélectionnées pour déterminer l'étendue de l'erreur d'échantillonnage dans l'estimation (voir la définition de l'erreur d'échantillonnage au tableau 3.1 et voir le chapitre 7 et l'annexe I pour une discussion plus détaillée). Les facteurs qui affectent l'ampleur de la variance d'échantillonnage sont notamment l'hétérogénéité de la variable à l'examen, la taille de l'échantillon et la conception de l'échantillon (ces aspects sont discutés dans les différentes sections du présent chapitre et au chapitre 7, et les principes fondamentaux de l'échantillonnage sont exposés à l'annexe I).

5. Ce chapitre et le suivant analysent en détail chacune des caractéristiques à prendre en considération pour concevoir un échantillon approprié en vue d'une enquête sur les ménages. D'une manière générale, l'accent est mis sur les enquêtes nationales, bien que toutes les techniques décrites ici puissent être appliquées à de vastes enquêtes infranationales comme celles qui portent sur une ou plusieurs régions, provinces, districts ou villes. Du fait de l'importance capitale que revêtent les

cadres d'échantillonnage pour une conception appropriée des échantillons, le chapitre 4 est entièrement consacré à cette question.

3.1.2. Glossaire des termes d'échantillonnage et des termes connexes

6. Nous commencerons par un glossaire des termes employés dans le présent chapitre et le chapitre suivant (voir le tableau 3.1). Ce glossaire n'a pas pour but de donner des définitions formelles des termes d'échantillonnage, dont certains sont mathématiques, mais plutôt de décrire comment ces termes sont employés dans le présent guide, en mettant évidemment l'accent sur leur utilisation dans le contexte des enquêtes sur les ménages.

Tableau 3.1

Glossaire des termes d'échantillonnage et des termes connexes

TERME	USAGE
Autopondération	Échantillon conçu de sorte que tous les cas aient la même pondération aux fins de l'enquête
Cadre(s) d'échantillonnage	Série d'informations sur la base desquelles l'échantillon est effectivement sélectionné, par exemple une liste ou une série de régions géographiques, soit un ensemble d'unités de population
Comptage rapide	Opération de mise à jour visant à mesurer la taille d'un échantillon au moyen d'un comptage approximatif des logements; voir également *quadrillage*
Corrélation intraclasse	Le coefficient de corrélation intraclasse mesure la ressemblance (homogénéité) des éléments
Dimensions de l'échantillon	Nombre de ménages ou de personnes sélectionnés
Domaine	Unités géographiques pour lesquelles doivent être fournies des estimations distinctes
EAS	Échantillon aléatoire simple (rarement utilisé dans les enquêtes sur les ménages)
Échantillon de conception complexe	Utilisation pour une enquête sur les ménages d'échantillons à phases multiples, d'échantillons en grappes et d'échantillons stratifiés, par opposition à des échantillons aléatoires simples
Échantillon-maître	« Super » échantillon devant être utilisé pour des enquêtes multiples et/ou plusieurs séries de la même enquête, habituellement à intervalles de dix ans
Échantillonnage en grappes	Échantillonnage dont l'avant-dernière phase porte sur une unité géographiquement définie, comme une zone d'énumération du recensement
Échantillonnage Epsem	Échantillonnage à égale probabilité
Échantillonnage géographique	Sélection des unités géographiques qui constituent un cadre d'échantillonnage (peut comprendre une sélection de segments, définis comme étant des subdivisions établies de circonscriptions administratives)
Échantillonnage non probabiliste	Voir dans la section 3.2.2 les descriptions de différents exemples de cette méthode : échantillonnage par quotas, échantillonnage sélectif, échantillonnage de commodité, échantillonnage par marche aléatoire

TERME	USAGE
Échantillonnage par étapes	Méthode consistant à choisir un échantillon de circonscriptions administratives et de ménages/personnes par étapes successives pour identifier les régions géographiques où l'enquête sera menée
Échantillonnage par phases, également appelé double échantillonnage ou échantillonnage post-stratifié	Sélection de l'échantillon sur (généralement) deux périodes, l'échantillon secondaire étant habituellement un sous-échantillon de l'échantillon primaire; à ne pas confondre avec l'échantillonnage tendanciel
Échantillonnage probabiliste	Méthode de sélection selon laquelle chaque unité de la population (personne, ménage, etc.) a des chances égales et connues d'inclusion dans l'échantillon
Échantillonnage selon probabilité proportionnelle à la taille	Sélection d'échantillons primaires, secondaires, etc., chacun étant choisi selon une probabilité proportionnelle à sa taille; voir également, dans le texte, échantillonnage selon probabilité proportionnelle à la taille estimative
Échantillonnage stratifié	Technique consistant à organiser le cadre d'échantillonnage en sous-groupes homogènes au plan interne et hétérogènes au plan externe pour veiller à ce que la sélection des échantillons soit « éparpillée » comme il convient parmi d'importants sous-groupes de population
Échantillonnage sur liste	Sélection sur une liste des unités qui constituent le cadre d'échantillonnage
Échantillonnage systématique	Sélection sur une liste, en commençant au hasard et à intervalles de sélection prédéterminés
Échantillonnage tendanciel	Échantillon conçu de manière à estimer le changement intervenu d'une période à l'autre
Effet de conception	Ratio de variance entre un échantillon de conception complexe et un échantillon aléatoire simple de mêmes dimensions; parfois appelé *effet de mise en grappes,* bien que l'effet de conception comprenne les effets non seulement de la mise en grappes mais aussi de la stratification
Erreur d'échantillonnage (erreur type)	Erreur qui caractérise l'estimation du fait que l'enquête porte sur un échantillon plutôt que sur l'ensemble de la population; racine carrée de la variance d'échantillonnage
Erreur type relative (coefficient de variation)	Erreur type en pourcentage de l'estimation issue de l'enquête, autrement dit erreur type divisée par estimation
Erreurs autres que les erreurs d'échantillonnage	Distorsion de l'estimation découlant d'erreurs de conception et d'exécution; ce terme se réfère à l'exactitude ou à la validité d'une estimation, par opposition à sa fiabilité ou à sa précision
Estimateur	Pour un échantillon de conception déterminée, l'estimateur est la méthode d'estimation du paramètre décrivant la population sur la base des données relatives à l'échantillon; par exemple, la moyenne arithmétique de l'échantillon est un estimateur
Exactitude (validité)	Voir erreur autre que d'échantillonnage
Fiabilité (précision, marge d'erreur)	Degré d'erreur d'échantillonnage que présente une estimation donnée
Fraction d'échantillonnage	Ratio entre la taille de l'échantillon et le nombre total d'unités de population
Grappe	Tendance des unités d'échantillonnage — personnes ou ménages — à présenter des caractéristiques similaires
Grappe compacte	Grappe composée de ménages géographiquement continus

TERME	USAGE
Grappe non compacte	Grappe composée de ménages géographiquement dispersés
Grappe, dimensions de la	Nombre (moyen) d'unités d'échantillonnage, personnes ou ménages, que comporte la grappe
Mesure de l'échantillon	Pour un échantillonnage en phases multiples, dénombrement ou estimation des dimensions (par exemple nombre de personnes) de chaque unité lors d'une phase déterminée
Niveau de confiance	Décrit le degré de confiance statistique qui caractérise la précision et la marge d'erreur des estimations, 95 % étant généralement considéré comme la norme
Phase de sélection à blanc	Pseudo-phase de sélection visant à simplifier la tâche manuelle consistant à identifier les sous-régions géographiques où se trouveront en définitive les grappes prises comme échantillon
Pondération	Inverse de la probabilité de sélection; facteur d'inflation appliqué aux données brutes; également appelée pondération de conception
Population cible	Définition de la population devant être couverte par l'enquête; également appelée univers de couverture
Quadrillage	Méthode consistant à « couvrir » les régions géographiques pour localiser les habitations et/ou les ménages, habituellement appliquée pour mettre à jour un cadre d'échantillonnage
Segment	Subdivision délimitée d'une grappe importante
Segmentation (morcellement)	Méthode habituellement utilisée sur le terrain consistant à subdiviser des grappes de taille inattendue pour alléger le travail d'établissement des listes
Stratification implicite	Méthode de stratification fondée sur un tri géographique du cadre d'échantillonnage et sur un échantillonnage systématique avec une probabilité proportionnelle aux dimensions de l'échantillon
Unité primaire d'échantillonnage (UPE)	Circonscription administrative géographiquement définie sélectionnée comme première phase de l'échantillonnage
Variance d'échantillonnage	Carré de l'erreur type ou erreur d'échantillonnage

3.1.3. Notations

7. Les notations utilisées dans le présent chapitre et les chapitres suivants du guide (voir le tableau 3.2) sont les notations standard. D'une manière générale, les majuscules dénotent les valeurs de la population et les minuscules les observations concernant l'échantillon. Par exemple, $\bar{y}$ représente des valeurs de la population, tandis que ^ est utilisé généralement pour dénoter des valeurs de l'échantillon. Il découle de ce qui précède que N représente les effectifs de la population et n la taille de l'échantillon. Il importe de noter que les paramètres de la population sont indiqués soit par des majuscules de l'alphabet latin, soit par des lettres grecques. Dans le tableau 3.2, par exemple, sont des valeurs de la population, tandis que sont généralement utilisés pour dénoter des valeurs de l'échantillon. Il découle de ce qui précède que N représente les effectifs de la population et la taille de l'échantillon. Il importe de noter que les paramètres de la population sont indiqués soit par des majuscules, soit par des lettres grecques. Par exemple, $\hat{\bar{Y}}$ et σ désignent respectivement les paramètres correspondant à la « moyenne de la population » et à la « déviation type ». Les estimateurs des paramètres de la population sont indiqués par le symbole ^ au-dessus du symbole correspondant au paramètre de la population.

Tableau 3.2

Notations sélectionnées utilisées pour les valeurs de la population et les caractéristiques des échantillons

	Notation		
		Échantillon	
Caractéristiques	Population	Estimations accompagnées du symbole ∧ au- dessus de la notation	Notation en minuscule
Unités	N	$\hat{n}$	n
Observations	$Y_1, Y_2, \ldots \ldots Y_i, \ldots \ldots Y_N$	$\hat{Y}_1, \hat{Y}_2, \ldots \ldots \hat{Y}_i, \ldots \ldots \hat{Y}_n$	$y_1, y_2, \ldots \ldots y_i, \ldots \ldots y_n$
Valeur moyenne	$\bar{Y}$	$\hat{\bar{Y}}$	$\bar{y}$
Proportion	P	$\hat{P}$	p
Estimateur du paramètre	θ	$\hat{\theta}$	t ou $\hat{\theta}$
Variance de *y*	$\sigma^2(y)$	$\hat{\sigma}^2(y)$	$s^2(y)$
Écart type de *y*	$\sigma(y)$	$\hat{\sigma}(y)$	$s(y)$
Corrélation intraclasse	δ	$\hat{\delta}$	$\hat{\delta}$
Ratio	R	$\hat{R}$	r
Somme	$\sum_{i=}^{N}$	$\sum_{i=}^{n}$	$\sum_{i=}^{n}$

3.2. Échantillon probabiliste et autres méthodes d'échantillonnage utilisées pour les enquêtes sur les ménages

8. Une discussion de la théorie des probabilités sortirait du cadre du présent guide, mais il importe d'expliquer pourquoi des méthodes de probabilité jouent un rôle indispensable dans la sélection des échantillons destinés aux enquêtes sur les ménages. L'on trouvera donc ci-après une définition et une description succinctes de l'échantillonnage probabiliste et une explication des raisons de son importance. D'autres méthodes, comme l'échantillonnage sélectif, l'échantillonnage par marche aléatoire, l'échantillonnage par quotas et l'échantillonnage de commodité, qui ne réunissent pas les conditions propres à l'échantillon probabiliste, sont mentionnées brièvement, avec une indication des raisons pour lesquelles elles ne sont pas recommandées aux fins des enquêtes sur les ménages.

3.2.1. Échantillonnage probabiliste

9. L'échantillonnage probabiliste, dans le contexte des enquêtes sur les ménages, désigne les moyens utilisés pour sélectionner les éléments de la population cible — unités géographiques, ménages et personnes — qui seront inclus dans l'enquête. Pour cela, il faut : *a*) que chaque élément ait une chance mathématique connue d'être sélectionné; *b*) que cette chance soit supérieure à zéro; et *c*) qu'elle soit

numériquement calculable. Il importe de noter que la chance que chaque élément a d'être sélectionné ne doit pas nécessairement être égale mais peut varier selon les objectifs de l'enquête.

10. C'est le caractère mathématique de l'échantillonnage probabiliste qui permet de tirer de l'enquête des estimations scientifiquement fondées. Un aspect plus important est cependant que c'est la base sur laquelle l'on peut déduire que les estimations relatives à l'échantillon correspondent à la population totale parmi laquelle l'échantillon a été sélectionné. Le fait que les erreurs d'échantillonnage peuvent être estimées au moyen des données rassemblées au sujet de l'échantillon est un sous-produit essentiel de l'échantillonnage probabiliste. Les méthodes d'échantillonnage non probabiliste ne présentent aucune de ces caractéristiques. De ce fait, il est vivement recommandé d'utiliser toujours un échantillonnage probabiliste aux fins des enquêtes sur les ménages, même si l'enquête peut coûter plus cher que s'il était utilisé des méthodes non scientifiques et non probabilistes.

3.2.1.1. Échantillonnage probabiliste par étapes

11. Comme indiqué ci-dessus, il y a lieu d'avoir recours à des méthodes d'échantillonnage probabiliste à chacune des étapes du processus de sélection de l'échantillon. Par exemple, la première étape de la sélection consiste généralement à choisir des unités géographiquement délimitées comme des villages, et la dernière à sélectionner les ménages ou personnes spécifiques à interroger. Ces deux étapes, de même que toutes les étapes intermédiaires, doivent reposer sur des méthodes d'échantillonnage probabiliste. L'on en trouvera ci-après un exemple.

> **Exemple**
>
> Prenons le cas d'un échantillon aléatoire simple (EAS) de 10 villages sélectionnés sur les 100 villages d'une province rurale. Supposons en outre que, pour chaque village faisant partie de l'échantillon, il est établi une liste complète des ménages. À partir de cette liste, il est sélectionné systématiquement un ménage sur cinq aux fins de l'enquête, quel que soit le nombre de ménages figurant sur la liste. Il s'agit d'un échantillon probabiliste avec une sélection en deux étapes et une probabilité de 10/100 pour la première étape et de 1/5 pour la seconde. La probabilité globale de sélection d'un ménage est de 1/50, c'est-à-dire de 10/100 multiplié par 1/5.

12. Bien qu'elle ne soit pas particulièrement efficace, la conception de cet échantillon illustre néanmoins comment les deux étapes de l'échantillonnage sont fondées sur les probabilités. Ainsi, les résultats de l'enquête peuvent être estimés *sans distorsion* en appliquant comme il convient les probabilités de sélection au stade de l'analyse des données (voir la discussion sur les informations concernant la pondération, au chapitre 6).

3.2.1.2. Calcul de la probabilité

13. Cet exemple illustre également comment les deux autres conditions qui caractérisent l'échantillonnage probabiliste ont été réunies. Premièrement, chaque village de la province a une chance *non zéro* d'être sélectionné. En revanche, si un ou plusieurs des villages avaient été écartés pour quelque raison que ce soit, par exemple des considérations de sécurité, la probabilité de sélection de ces villages aurait été égale à zéro et la nature probabiliste de l'échantillon aurait donc été violée. Dans l'exemple ci-dessus, les ménages ont également été sélectionnés avec une probabilité autre que zéro. Cependant, si certains d'entre eux avaient été délibérément exclus pour des raisons, par exemple, d'inaccessibilité, ils auraient eu une probabilité nulle et l'échantillon aurait alors été non probabiliste.

L'on verra dans la section 3.2.1.3 comment faire face à une telle situation lorsque certaines régions sont exclues de l'enquête.

14. Deuxièmement, la probabilité de sélection aussi bien des villages que des ménages pouvait effectivement être *calculée* sur la base des informations disponibles. Dans le cas des villages, tant la taille de l'échantillon (10) que l'effectif de la population (100) étaient connus, et tels étaient les paramètres qui ont défini la probabilité, 10/100. Pour les ménages, le calcul de la probabilité était légèrement différent car nous ne savions pas avant l'enquête combien de ménages allaient être sélectionnés dans chacun des villages faisant partie de l'échantillon. Il nous avait simplement été dit d'en sélectionner un sur cinq. Ainsi, s'il y avait eu en tout 100 ménages dans le village A et 75 dans le village B, nous en aurions sélectionné 20 et 15 respectivement. Il n'en demeure pas moins que la probabilité de sélection d'un ménage était de 1/5, quels que soient les effectifs de la population et la taille de l'échantillon (20/100 = 1/5, mais tel est aussi le cas de 15/75).

15. Toujours dans le cas de l'exemple susmentionné, la probabilité de sélection à la deuxième étape aurait pu être calculée pour vérification après la fin de l'enquête. Lorsque m_i et M_i sont connus et lorsque m_i et M_i sont, respectivement, le nombre de ménages figurant dans l'échantillon et le nombre total de ménages dans le $i^{ième}$ village, la probabilité serait égale à m_i/M_i. Il y aurait dix de ces probabilités, une pour chaque village. Comme on l'a vu, cependant, ce ratio serait toujours de 1/5 pour le type d'échantillon spécifié. Il serait donc superflu de rechercher le nombre d'échantillons et le nombre total de ménages uniquement pour calculer la probabilité secondaire. À des fins de *contrôle de la qualité,* il pourrait néanmoins être utile d'obtenir ces chiffres pour s'assurer que le ratio 1/5 a été respecté.

3.2.1.3. Cas d'une population cible mal définie

16. Les conditions que doit réunir un échantillonnage probabiliste ne sont parfois pas respectées parce que la *population cible* que l'enquête est censée couvrir a été définie sur la base de critères excessivement vagues. Par exemple, la population cible idéale pourrait être tous les ménages du pays. Cependant, lorsque l'enquête est conçue et/ou exécutée, il arrive fréquemment que certains sous-groupes de population comme les ménages nomades, les bateliers et les populations qui vivent dans des régions accidentées et par conséquent inaccessibles soient délibérément exclus. Dans d'autres cas, une population cible qui constitue un groupe spécial restreint, comme les femmes ayant été mariées au moins une fois dans leur vie ou les jeunes de moins de 25 ans, exclut d'importants sous-groupes. Par exemple, une population cible définie comme les jeunes de moins de 25 ans peut exclure les jeunes sous les drapeaux ou les jeunes détenus, ou les jeunes vivant, pour d'autres raisons, dans des institutions publiques.

17. Si la population effectivement couverte par l'enquête s'écarte de la population cible initialement prévue, les enquêteurs doivent veiller à définir la population cible avec plus de précision. Cela est important non seulement pour que les résultats de l'enquête soient plus clairs pour les usagers mais aussi pour que les conditions que suppose un échantillonnage probabiliste soient réunies. Dans l'exemple susmentionné des jeunes de moins de 25 ans, la population cible doit être décrite de manière plus précise et redéfinie comme les *jeunes civils de moins de 25 ans ne vivant pas en institution*. Autrement, la couverture de l'enquête doit être élargie de manière à englober les sous-groupes exclus.

18. Il importe par conséquent de définir très soigneusement la population cible de manière à ne couvrir que ceux de ses membres qui auront effectivement *la possibilité d'être sélectionnés* pour l'en-

quête. Lorsque certains sous-groupes sont délibérément exclus, il importe d'appliquer des méthodes probabilistes à la population qui constitue effectivement le cadre de l'enquête. Il faut également que les responsables de l'enquête décrivent clairement aux usagers, lorsque les résultats de l'enquête sont publiés, quels sont les segments de la population qui ont été inclus ou au contraire exclus.

3.2.2. Méthodes d'échantillonnage non probabiliste

19. Il n'existe pas de théorie statistique, comme pour l'échantillonnage probabiliste, dont on puisse s'inspirer en ce qui concerne le recours à des échantillons non probabilistes. Ces échantillons ne peuvent être évalués qu'au moyen d'une appréciation subjective. Si les techniques probabilistes ne sont pas utilisées, par conséquent, les estimations résultant de l'enquête seront faussées. En outre, l'ampleur de ces distorsions et, souvent, l'occasion de savoir si elles entraînent une sous-estimation ou plutôt une surestimation seront inconnues. Comme on l'a déjà dit, la précision des estimations, c'est-à-dire leurs erreurs types, peut être estimée, lorsque l'on utilise un échantillonnage probabiliste. Ce n'est qu'ainsi qu'il est possible d'évaluer la fiabilité des estimations issues de l'enquête et de construire des intervalles de confiance autour des estimations. Des estimations subjectives peuvent également être faites dans certains cas au moyen d'un échantillonnage probabiliste, par exemple si l'on veut que la répartition de la population couverte par l'enquête soit conforme à d'autres critères (voir le chapitre 6 pour une discussion plus approfondie de cette question).

20. En dépit de leurs déficiences théoriques, des échantillons non probabilistes sont fréquemment utilisés dans différents contextes et dans diverses situations. La justification offerte par les praticiens est généralement fondée sur des considérations de coûts ou de commodité ou même sur la crainte des enquêteurs qu'un échantillon « aléatoire » ne représente pas comme il convient la population cible. Dans le contexte des enquêtes sur les ménages, nous examinerons brièvement plusieurs types de méthodes non probabilistes, essentiellement au moyen d'exemples, et indiquerons certaines des raisons pour lesquelles elles ne doivent pas être utilisées.

3.2.2.1. Échantillonnage sélectif

21. L'échantillonnage sélectif est une méthode qui fait appel à des « experts » pour choisir les éléments à prendre comme échantillon. Les partisans de cette méthode font valoir que cette méthode élimine le risque, lorsque l'on utilise les techniques aléatoires, de sélection d'un échantillon sortant de l'ordinaire ou d'un « mauvais » échantillon, par exemple lorsque tous les éléments de l'échantillon, par malchance, se trouvent être réunis dans la région nord-ouest.

Exemple

Un exemple d'échantillonnage sélectif dans le contexte d'une enquête sur les ménages serait celui d'un groupe d'experts qui choisissent délibérément les circonscriptions géographiques à utiliser comme éléments lors de la sélection primaire et qui ont fondé leur décision sur les districts qui apparaissent comme typiques ou représentatifs dans un certain sens ou dans un certain contexte.

22. La principale difficulté que soulève ce type d'échantillonnage tient au caractère subjectif de la décision prise sur le point de savoir ce qui constitue une série représentative de districts. Paradoxalement, le choix dépend directement aussi du *choix* des experts eux-mêmes. Avec un échantillonnage probabiliste, en revanche, les districts seraient stratifiés en utilisant, si besoin est, les critères retenus par les concepteurs. Il y a lieu de noter que les critères de stratification peuvent même être *subjectifs,*

bien qu'il existe des directives pour que soient appliqués des critères plus objectifs (voir la section 3.4 concernant la stratification). En pareil cas, un échantillon probabiliste de districts (sélectionnés selon une méthode parmi bien d'autres) sera choisi *dans chaque strate*. Il y a lieu de noter que la stratification réduit considérablement le risque de sélection d'un échantillon sortant de l'ordinaire comme celui qui a été évoqué ci-dessus. *Telle est précisément la raison pour laquelle la stratification a été inventée*. Avec un exemple stratifié, chaque district aura une chance de sélection connue, autre que zéro, qui n'est pas faussée ni affectée par une opinion subjective (même lorsque les strates elles-mêmes sont définies de façon subjective). D'un autre côté, l'échantillonnage sélectif n'a aucun mécanisme permettant de garantir que chaque district ait une chance d'inclusion autre que zéro ni de calculer la probabilité de sélection de ceux qui seront retenus en définitive.

3.2.2.2. Échantillonnage par marche aléatoire et par quotas

23. Un autre type d'échantillonnage non probabiliste qui est largement utilisé est la méthode dite de la marche aléatoire, appliquée pendant la dernière étape d'une enquête sur les ménages. Cette technique est fréquemment utilisée même si les éléments de l'échantillon ont été sélectionnés lors des étapes antérieures par des méthodes probabilistes légitimes. L'exemple ci-après illustre un type d'échantillonnage qui est une combinaison d'échantillonnage par marche aléatoire et par quotas. Cette dernière méthode est une autre technique non probabiliste consistant à imposer aux enquêteurs des quotas de certains types de personnes à interroger.

Exemple

Selon cette méthode, il serait donné pour instruction aux enquêteurs de commencer le processus d'entrevue en un point géographique choisi au hasard, par exemple dans un village, et de suivre ensuite un itinéraire déterminé pour sélectionner les ménages à interroger, par exemple en sélectionnant chaque $n^{ième}$ ménage ou en déterminant, tout au long de l'itinéraire suivi, si les ménages comportent des membres d'une population cible particulière, comme des enfants de moins de 5 ans. Dans ce dernier cas, chacun de ces ménages serait interrogé pour l'enquête jusqu'à ce qu'un quota prédéterminé soit atteint.

24. Cette méthodologie est fréquemment présentée comme un moyen d'éviter le processus, long et coûteux, consistant à dénombrer préalablement tous les ménages faisant partie de l'échantillon — village, grappe ou segment — avant de sélectionner ceux qui seront interrogés. L'on a fait valoir également que cette méthode permet d'éviter les non-réponses étant donné que l'enquêteur continue d'interroger les ménages jusqu'à ce qu'il obtienne suffisamment de réponses pour atteindre le quota fixé. En outre, les partisans de cette méthode soutiennent que cette technique évite les distorsions aussi longtemps que le point de départ, le long du cheminement suivi, est déterminé au hasard. Ils soutiennent en outre que les probabilités de sélection peuvent être calculées comme il convient comme étant le nombre de ménages sélectionnés divisé par le nombre de ménages que comporte le village, si ce dernier chiffre est connu ou peut être estimé de façon assez précise.

25. Étant donné les conditions indiquées ci-dessus, un échantillon probabiliste est théoriquement possible. Dans la pratique, cependant, il est douteux que cela ait jamais été le cas. Cette approche échoue habituellement en raison : *a*) du comportement de l'enquêteur; et *b*) du traitement des ménages qui ne répondent pas, y compris ceux qui peuvent s'abstenir de répondre. D'innombrables études ont établi que lorsque les enquêteurs se voient donner le pouvoir de sélectionner l'échantillon sur le terrain, il en résulte des distorsions. Par exemple, la taille moyenne (nombre de personnes) des mé-

nages pris comme échantillon est habituellement plus réduite que celle de l'ensemble des ménages[1]. Il est en effet naturel pour un enquêteur d'éviter un ménage qui peut apparaître comme pouvant susciter des difficultés. Il est par conséquent plus simple de sauter un ménage dont le chien paraît être méchant ou un ménage dont l'accès est difficile et de privilégier plutôt les ménages voisins qui ne soulèvent pas de tels problèmes.

26. En remplaçant les ménages qui ne répondent pas par des ménages qui répondent, il est introduit une distorsion qui privilégie les ménages disposés à coopérer et aisément accessibles. Il y a manifestement des différences dans les caractéristiques des ménages, selon qu'ils sont disponibles et selon qu'ils sont disposés à participer à l'enquête. Avec la méthode de l'échantillonnage par quotas, les personnes qu'il est difficile de contacter ou qui hésitent à participer à l'enquête risquent davantage d'être sous-représentées que dans le cas d'un échantillon probabiliste. Dans ce dernier cas, les enquêteurs sont généralement tenus de revenir à la charge plusieurs fois pour interroger les ménages dont les membres sont temporairement indisponibles. De plus, dans le cas des enquêtes basées sur des méthodes probabilistes, les enquêteurs reçoivent généralement pour instruction de s'efforcer tout particulièrement de convaincre les ménages hésitants de participer à l'enquête.

3.2.2.3. Échantillon de commodité

27. L'échantillon de commodité est une méthode qui est elle aussi largement utilisée car son application est très simple. Bien que cette méthode soit rarement utilisée dans le contexte des enquêtes sur les ménages, elle l'est dans de nombreux cas, par exemple pour une enquête auprès des élèves d'un échantillon d'écoles délibérément choisies parce qu'elles sont aisément accessibles et parce que l'on sait qu'elles sont disposées à coopérer, autrement dit parce qu'elles sont commodes. Une autre méthode actuellement en vogue consiste à administrer un sondage d'opinion instantané sur différents sites Internet. La raison pour laquelle des échantillons de ce type introduisent inévitablement des distorsions et ne doivent pas être utilisés pour en tirer des déductions concernant l'ensemble de la population est peut-être évidente.

3.3. Détermination de la taille de l'échantillon pour une enquête sur les ménages

28. La présente section est extrêmement détaillée en raison de l'importance que revêt la taille de l'échantillon pour l'ensemble de l'opération et pour le coût de l'enquête. Elle est importante car elle affecte non seulement le nombre de ménages à interroger mais aussi le nombre de régions géographiques qui constitueront des unités primaires d'échantillonnage (UPE), le nombre d'enquêteurs à recruter, la charge de travail de chaque enquêteur, etc. Les facteurs et paramètres à prendre en considération pour déterminer la taille de l'échantillon sont nombreux mais sont liés essentiellement aux objectifs de mesure de l'enquête. Nous discuterons de la détermination de la taille de l'échantillon à la lumière des principales estimations à établir, des populations cibles, du nombre de ménages devant être inclus dans l'échantillon pour atteindre les populations cibles voulues, de la précision et du

1 Beaucoup d'organisations ont souvent eu pour pratique de veiller à ce que la désignation des ménages sélectionnés pour faire partie de l'échantillon soit effectuée au bureau, où l'opération peut être supervisée de plus près. De plus, l'échantillon doit être sélectionné par quelqu'un qui soit n'a pas participé à l'établissement de la liste des ménages avant la sélection de l'échantillon, soit n'est pas familiarisé avec la situation qui prévaut effectivement sur le terrain.

degré de confiance désirés, des domaines d'estimation, de la question de savoir si la mesure porte sur des chiffres absolus ou un changement, de l'effet de grappe, de l'élément non-réponse à prendre en considération et le budget disponible. Manifestement, la taille de l'échantillon est l'élément central qui influe sur toute la conception de l'échantillon.

3.3.1. Ordre de grandeur des estimations

29. Dans les enquêtes sur les ménages, qu'elles soient de portée générale ou soient axées sur un sujet spécifique comme la santé ou l'activité économique, toutes les estimations (souvent appelées *indicateurs*) à générer sur la base des résultats de l'enquête exigent, pour être fiables, un échantillon d'une taille différente. La taille de l'échantillon dépend de l'ordre de grandeur de l'estimation, c'est-à-dire de la proportion que celle-ci représentera par rapport à la population totale. Par exemple, pour estimer de façon fiable la proportion des ménages ayant accès à l'eau salubre, il faut prendre un échantillon d'une taille différente de celle de l'échantillon à constituer pour estimer la proportion d'adultes qui ne travaillent pas au moment considéré.

30. Les expressions à utiliser pour calculer la taille de l'échantillon sont fondées sur des théorèmes probabilistes, à savoir que le paramètre réel d'une population doit se trouver à l'intérieur d'un intervalle avec une probabilité donnée (niveau de confiance). La largeur de l'intervalle (ou précision) dépend de la variance de la population visée au tableau 3.2, du degré de confiance et de la taille de l'échantillon. D'une manière générale, plus l'hétérogénéité de la population ou plus le degré de confiance souhaité sont grands, et plus l'intervalle doit être large. D'un autre côté, la largeur de l'intervalle diminuera à mesure que la taille de l'échantillon augmente. Des exemples d'intervalles de confiance sont donnés au paragraphe 22 du chapitre 7. L'équation ci-après représente un intervalle de confiance de la moyenne de la population $\bar{Y}$ compte tenu de l'estimation de la moyenne de la population $\hat{\bar{Y}}$ sur la base d'un échantillon aléatoire simple sans remplacement de dimension n.

$$P\left[\hat{\bar{Y}} - z_{1-\alpha}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2(y)}{n}} \le \bar{Y} \le \hat{\bar{Y}} + z_{1-\alpha}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2(y)}{n}}\right] = (1-\alpha)100\ \% \qquad (3.5)$$

où $1-\alpha$ est le coefficient de confiance pour l'intervalle. Il y a lieu de noter en outre que lorsque l'on prend en considération l'estimation d'une proportion p, $\sigma^2(y) = p\,(1-p)$.

31. Dans la pratique, l'enquête elle-même ne peut reposer que sur un échantillon de même taille. Pour calculer la taille de l'échantillon, il faut choisir parmi les nombreuses estimations qui doivent être mesurées lors de l'enquête. Par exemple, si l'estimation clé est le taux de chômage, c'est sur cette base que serait calculée la taille de l'échantillon[2]. Lorsque les indicateurs clés sont nombreux, une convention parfois appliquée consiste à calculer la taille de l'échantillon requis pour chaque indicateur et de retenir celui qui exige l'échantillon de plus grande taille. Généralement, c'est l'indicateur pour lequel la population de base est une « population cible secondaire » la plus réduite pour ce qui est de sa proportion par rapport à la population totale. Il faut évidemment tenir compte du degré de

2 Il est quelque peu paradoxal qu'il faille, pour calculer la taille de l'échantillon, savoir quelle est la valeur approximative de l'estimation à mesurer. Cette valeur peut cependant être « devinée » de différentes façons, par exemple en utilisant les données provenant d'un recensement, d'une enquête semblable, d'un pays voisin, d'une enquête pilote, etc.

précision requis (voir ci-après). Lorsque la taille de l'échantillon est fondée sur une telle estimation, chacune des autres estimations clés sera mesurée avec la même fiabilité ou une fiabilité plus grande.

32. À défaut, la taille de l'échantillon peut être calculée sur la base d'une proportion relativement restreinte de la population cible plutôt que d'un indicateur déterminé. Telle est généralement l'approche idéale pour une enquête sur les ménages de caractère général portant sur plusieurs questions hétéroclites, auquel cas il risque d'être difficile et imprudent de fonder la taille de l'échantillon sur un indicateur qui concerne un seul aspect. Les responsables de l'enquête pourront par conséquent décider de calculer la taille de l'échantillon en fonction de la possibilité de mesurer de façon fiable une caractéristique que possèdent 5 % ou 10 % de la population, le choix dépendant de considérations budgétaires.

3.3.2. Population cible

33. La taille de l'échantillon dépend également de la population cible que devra couvrir l'enquête. Comme dans le cas des indicateurs, une enquête sur les ménages porte fréquemment sur plusieurs populations cibles. Une enquête sur la santé, par exemple, pourra viser : *a*) les ménages, pour évaluer l'accès à l'eau salubre et à l'assainissement; tout en ciblant aussi *b*) toutes les personnes, pour estimer des situations chroniques et aiguës; *c*) les femmes de 14 à 49 ans pour identifier les indicateurs de la santé génésique; et *d*) les enfants de moins de 5 ans pour obtenir des mesures anthropométriques du poids et de la taille.

34. Pour calculer la taille de l'échantillon, il faut par conséquent prendre en considération chacune des populations cibles. Comme indiqué ci-dessus, les enquêtes sur les ménages portent fréquemment sur de multiples populations cibles dont chacune présente le même intérêt pour ce qui est des mesures à opérer. Encore une fois, il est possible de centrer l'enquête sur la population cible la plus réduite pour déterminer la taille de l'échantillon. Par exemple, si les enfants de moins de 5 ans constituent un groupe cible important pour l'enquête, c'est en fonction de ce groupe que devra être déterminée la taille de l'échantillon. En utilisant le concept décrit au paragraphe 32, les responsables de l'enquête pourront décider de calculer la taille de l'échantillon de manière à estimer l'une des caractéristiques de 10 % des enfants de moins de 5 ans. La taille de l'échantillon correspondant sera bien plus grande que celle de l'échantillon à retenir pour un groupe cible comprenant toutes les personnes ou tous les ménages.

3.3.3. Précision et confiance statistique

35. Il a été suggéré plus haut que les estimations, surtout celles qui concernent les indicateurs clés, doivent être *fiables*. La taille de l'échantillon dépend directement du degré de précision que doivent présenter les indicateurs. Plus les estimations doivent être précises ou fiables, et plus l'échantillon devra être nombreux, et ce de plusieurs ordres de grandeur. Si l'on veut que les estimations soient deux fois plus fiables, par exemple, il pourra être nécessaire de *quadrupler* la taille de l'échantillon. Les responsables de l'enquête devront manifestement être conscients de l'impact qu'une précision excessive aura sur la taille de l'échantillon et par conséquent sur le coût de l'enquête. Inversement, ils devront veiller à ne pas choisir un échantillon de taille si réduite que les principaux indicateurs seront trop peu fiables pour être utiles pour l'analyse ou la planification.

36. De même, la taille de l'échantillon augmente parallèlement au degré de confiance requis pour maintenir les précisions données. Un niveau de confiance de 95 % est presque universellement consi-

déré comme la norme, et la taille de l'échantillon à retenir pour l'obtenir est calculée en conséquence (voir le paragraphe 30 ci-dessus).

37. Compte tenu des indicateurs, une convention consiste, dans nombre d'enquêtes bien conçues, à utiliser comme norme de précision une marge d'erreur *relative* de 10 % au niveau de confiance de 95 % concernant les principaux indicateurs à estimer, ce qui signifie que l'erreur type d'un indicateur clé ne doit pas dépasser 5 % de l'estimation elle-même. Cela est calculé selon la formule (2 * 0,05*x,* où *x* est l'estimation). Par exemple, si la proportion estimative de personnes que compte la population active est de 65 %, l'erreur type ne devra pas dépasser 3,25 %, c'est-à-dire 0,65 multiplié par 0,05. Deux fois 0,0325, soit 0,065, est la marge d'erreur relative au niveau de confiance de 95 %. Par exemple, au paragraphe 30 précédent, nous avons :

$$\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2(y)}{n}} = 0{,}05 * \hat{\bar{Y}} \qquad (3.6)$$

38. La taille de l'échantillon nécessaire pour s'en tenir à une marge d'erreur relative de 10 % est donc le quart de la taille de l'échantillon qui doit être retenu lorsque la marge d'erreur relative est fixée à 5 %. Une marge d'erreur relative de 20 % est généralement considérée comme le maximum tolérable pour les indicateurs importants (bien que nous ne la recommandions pas). En effet, l'intervalle de confiance qui entoure les estimations lorsque l'erreur tolérée est plus grande est trop large pour pouvoir obtenir des résultats utiles pour la plupart des analyses ou pour la formulation des politiques. D'une manière générale, si le budget le permet, nous recommandons des marges d'erreur relative de 5 à 10 % pour les principaux indicateurs.

3.3.4. Groupes d'analyse : domaines

39. Un autre élément important qui influe directement sur la taille de l'échantillon est le nombre de domaines. Les domaines sont généralement définis comme étant les sous-groupes d'analyse pour lesquels l'on souhaite obtenir des données *également* fiables. La taille de l'échantillon se trouve accrue, approximativement[3], par un facteur égal au nombre de domaines visés. Cela, cependant, n'est vrai que si chacun des domaines est caractérisé par une variabilité semblable (voir la note 3 pour un éclaircissement). En effet, la taille de l'échantillon à retenir pour un niveau de précision donné ne dépend pas de la taille de la population elle-même, sauf lorsque l'échantillon représente un pourcentage significatif, par exemple 5 % au moins de la population (ce qui est rarement le cas des enquêtes sur les ménages). Ainsi, la taille de l'échantillon à sélectionner pour une seule province (si l'enquête ne doit effectivement porter que sur une seule province) serait identique à celle de l'échantillon à sélectionner pour l'ensemble du pays. Il s'agit là d'un point extrêmement important qui est souvent mal compris par les praticiens, lesquels pensent à tort que, plus la population est nombreuse, et plus l'échantillon devra lui aussi être nombreux.

40. Ainsi, lorsque l'on cherche à rassembler des données uniquement au plan national, il y a un seul domaine, et la taille de l'échantillon à retenir s'applique par conséquent à l'échantillon utilisé dans l'ensemble du pays. S'il a cependant été décidé qu'il faut obtenir des résultats également fiables pour les ménages vivant en milieu urbain et en milieu rural, séparément, il faudra calculer la taille

3 Tel est le cas lorsque l'on vise le même degré de fiabilité pour chacun des domaines.

de l'échantillon à retenir pour chaque domaine pour essayer d'obtenir le résultat final. D'une manière générale, la taille de l'échantillon à utiliser pour chacun des domaines pertinents devra être calculé de sorte que, s'il y a D_1, D_2,... ... D_k domaines, il y aura nécessairement au moins n_1, n_2, ... n_k tailles d'échantillon, qui dépendront de la variabilité de la caractéristique à mesurer dans chacun des domaines ainsi que du degré de confiance et de précision requis. La taille totale de l'échantillon à utiliser pour l'enquête sera par conséquent $n = n_1, + n_2, + ... + n_k$.

3.3.4.1. Sur-échantillonnage pour les estimations de domaines

41. Exiger une fiabilité égale pour les différents domaines a pour importante conséquence qu'il faut utiliser des taux d'échantillonnage disproportionnés. Ainsi, lorsque la répartition n'est pas égale à 50-50, comme ce sera généralement le cas pour les domaines urbain et rural, un sur-échantillonnage délibéré du secteur urbain, par exemple, sera habituellement nécessaire dans la plupart des pays pour obtenir une fiabilité égale. Il importe néanmoins de souligner que le sur-échantillonnage, dans un domaine déterminé d'une enquête nationale, est dicté principalement par la nécessité d'obtenir des estimations caractérisées par un niveau de confiance spécifique.

42. Il importe de noter également deux conséquences d'un échantillonnage délibéré de ce groupe, qu'il s'agisse de domaines ou de strates. Premièrement, il faut utiliser lors de l'enquête des pondérations compensatoires pour obtenir des estimations valables au plan national. Deuxièmement et surtout, les estimations nationales seront un peu moins fiables qu'elles le seraient si l'échantillon était proportionnellement réparti entre les divers sous-groupes.

3.3.4.2. Choix des domaines

43. Les sous-régions géographiques sont évidemment importantes et l'on a toujours tendance à les considérer comme des domaines aux fins des estimations. Dans une enquête d'envergure nationale, par exemple, les usagers veulent souvent des données non seulement pour chacune des principales régions mais aussi pour chaque province. Manifestement, le nombre de domaines doit être judicieusement choisi et tel est le cas aussi des groupes devant constituer ces domaines. Une stratégie plausible peut consister à déterminer quels sont les groupes d'estimations qui, en dépit de leur importance, n'exigent pas une fiabilité *égale* lors de la mesure. Les groupes d'estimations doivent être plutôt considérés aux fins de l'analyse comme les grandes *catégories* de tabulation, plutôt que comme des domaines. La taille des échantillons à retenir dans chaque cas particulier sera donc beaucoup plus réduite que si les groupes d'estimations étaient considérés comme des domaines, de sorte que leur fiabilité serait moindre aussi. Il y a lieu de noter cependant que le sur-échantillonnage dans un domaine spécifique peut être justifié par la nécessité d'obtenir dans ce domaine des estimations caractérisées par un degré de confiance et de précision donné, indépendamment des estimations établies au plan national.

Exemple

L'exemple ci-après montre comment il serait procédé à l'échantillonnage et quel serait son impact sur la fiabilité des données si les milieux urbain et rural étaient considérés comme des groupes de tabulation plutôt que comme des domaines. Supposons que la population soit rurale à concurrence de 60 % et urbaine à concurrence de 40 %. Si, pour atteindre un degré de précision spécifié, la taille de l'échantillon a été déterminée comme devant être de 8 000 ménages, il faudra prendre un échantillon de 16 000 ménages si le milieu urbain et le milieu rural

étaient considérés comme domaines distincts, c'est-à-dire 8 000 ménages dans chaque secteur. Au contraire, si les milieux urbain et rural sont considérés comme des groupes de tabulation, il serait sélectionné un échantillon national de 8 000 ménages, répartis proportionnellement entre les ménages ruraux et les ménages urbains, ce qui donnerait 4 800 et 3 200 ménages respectivement. Supposons en outre que l'erreur type prévue pour une caractéristique de 10 %, sur la base de l'échantillon de 8 000 ménages, est de 0,7 %. Telle est l'erreur type qui s'applique à l'estimation nationale (ou aux milieux urbain et rural séparément s'il est pris un échantillon de 8 000 ménages dans chaque domaine). Pour un échantillon national de 8 000 ménages sélectionnés proportionnellement sur une base urbaine et rurale, l'erreur type correspondante serait d'environ 0,9 %, c'est-à-dire le produit de la racine carrée du ratio des tailles de l'échantillon et de l'erreur type de l'estimation nationale, soit :

$$\left(\left[\sqrt{8\,000 / 4\,800}\right] \times 0{,}7\right).$$

Pour les ménages urbains, l'erreur type serait d'environ 1,1 % ou

$$\left(\left[\sqrt{4\,800 / 3\,200}\right] \times 0{,}7\right).$$

Une autre façon d'évaluer cet effet est de considérer que les erreurs types pour toutes les estimations rurales dépasseraient de 29 % environ

$$\left(\sqrt{8\,000 / 4\,800}\right),$$

celles des estimations nationales; pour les ménages urbains, la différence serait d'environ 58 % :

$$\left(\sqrt{8\,000 / 3\,200}\right).$$

44. Il y a lieu de noter que la dernière phrase de cet exemple s'applique sans égard à ce qu'est l'erreur type nationale. Autrement dit, elle s'applique à toutes les estimations établies lors de l'enquête. Il est donc possible d'analyser, avant l'échantillonnage, l'impact sur la fiabilité des estimations concernant les divers sous-groupes pouvant être considérés comme des domaines. Ainsi, les enquêteurs disposeraient d'informations qui les aideraient à déterminer si les domaines potentiels doivent être considérés comme des groupes de tabulation. Comme indiqué ci-dessus, cela signifie qu'il serait procédé à une allocation proportionnelle plutôt qu'à une allocation égale de l'échantillon. Par exemple, s'il est prévu une enquête nationale pour un pays dont 20 % de la population seulement est urbaine, la taille de l'échantillon urbain ne représenterait que 20 % de la taille de l'échantillon total. Ainsi, l'erreur d'échantillonnage pour les estimations urbaines serait deux fois plus élevée (racine carrée de 0,8 n/0,2 n) que pour les estimations rurales et deux fois un quart plus élevée que pour les estimations nationales (racine carrée de n/0,2 n). En pareil cas, les responsables de l'enquête pourront déterminer qu'un sur-échantillonnage s'impose dans le secteur urbain[4], ce qui aura pour effet, dans la pratique, de créer des domaines urbain et rural distincts.

45. De même, l'on peut analyser la corrélation entre les erreurs types et les domaines par opposition aux groupes de tabulation pour déterminer s'il y a lieu d'utiliser des régions ou d'autres unités

4 Tel serait le cas par exemple si les erreurs types relatives prévues pour l'un quelconque des principaux indicateurs urbains dépassaient, par exemple, 7,5 % (soit 15 % à un niveau de confiance de 95 %, c'est-à-dire le maximum tolérable suggéré dans le présent guide).

géographiques infranationales comme domaines et, dans l'affirmative, combien. Comme, dans le cas des domaines, les échantillons doivent être de même taille, l'utilisation de dix régions exigerait un échantillon représentant dix fois la taille de l'échantillon national, mais ce chiffre se trouverait réduit de moitié si l'on pouvait se contenter de cinq régions. De même, si les régions sont plutôt considérées comme des groupes de tabulation, l'échantillon national serait réparti proportionnellement entre elles. En pareil cas, la région *moyenne* serait caractérisée par des erreurs types 3,2 fois supérieures environ aux estimations nationales s'il y avait dix régions, mais elle serait seulement du double s'il n'y en avait que cinq.

3.3.5. Effets dus à la mise en grappes

46. La présente section aborde la question de savoir comment la détermination de la taille de l'échantillon est affectée (une discussion plus détaillée sur l'échantillonnage en grappes figure à la section 3.5). La mesure dans laquelle un échantillon est *mis en grappes* affecte la fiabilité et la précision des estimations et par conséquent la taille de l'échantillon. Dans le cas des enquêtes sur les ménages, les effets dus à la mise en grappes sont imputables : *a*) aux avant-dernières unités d'échantillonnage, généralement appelées « grappes », qui peuvent être des ménages ou des pâtés de maisons; *b*) aux ménages pris comme échantillon; *c*) à la taille et/ou à la variabilité des grappes; et *d*) à la méthode d'échantillonnage des ménages des grappes sélectionnées. Les effets de la mise en grappes ainsi que de la stratification peuvent être mesurés numériquement par l'effet de conception, ou *deff,* qui indique le surcroît de variance de l'échantillonnage (carré de l'erreur type) pour l'échantillon en grappes stratifié par rapport à un échantillon aléatoire simple de même taille. La stratification tend à *réduire* la variance de l'échantillonnage mais seulement dans une mesure réduite. En revanche, la mise en grappes accroît considérablement la variance. Par conséquent, le *deff* indique essentiellement l'étendue de la mise en grappes d'un échantillon.

47. Pour concevoir efficacement l'échantillon, il faut procéder à une mise en grappes pour maîtriser les coûts, mais il faut aussi réduire autant que possible l'effet dû à la conception pour que les résultats soient utiles et fiables. Regrettablement, le *deff* n'est pas connu avant le début de l'enquête et il ne peut être estimé qu'à posteriori, sur la base des données elles-mêmes. Lorsqu'il a déjà été réalisé des enquêtes, ou lorsque des enquêtes semblables ont été menées dans d'autres pays, les valeurs du *deff* tirées de ces enquêtes peuvent être utilisées pour estimer la taille de l'échantillon.

48. Pour réduire autant que cela est possible l'effet dû à la conception, celle-ci doit suivre les principes généraux ci-après (voir également les lignes directrices résumées à la fin du présent chapitre, p. 82), selon lesquels il faut :

a) Utiliser autant de grappes que possible;

b) Utiliser pour chaque grappe un nombre de ménages aussi réduit que possible;

c) Utiliser des grappes de taille constante plutôt que variable;

d) Sélectionner lors de la dernière étape un échantillon systématique de ménages géographiquement dispersés plutôt qu'un segment de ménages géographiquement contigus.

49. Ainsi, pour un échantillon de 12 000 ménages, il est préférable de sélectionner 600 grappes de 20 ménages chacune plutôt que 400 grappes de 30 ménages chacune. L'effet dû à la conception de l'échantillon est beaucoup plus réduit dans le premier cas. En outre, le *deff* est minimisé si les ménages sont choisis systématiquement parmi tous les ménages de la même grappe plutôt que sélec-

tionnés dans des sous-segments géographiques contigus. Lorsque ces règles sont suivies, l'effet dû à la conception est généralement assez réduit.

3.3.6. Ajustement de la taille de l'échantillon en prévision des non-réponses

50. Il est fréquent, dans les enquêtes, d'accroître la taille de l'échantillon dans une proportion correspondant au taux prévu de non-réponse, ce qui permet de faire en sorte que le nombre de personnes effectivement interrogées lors de l'enquête soit aussi proche que possible de la taille idéale de l'échantillon.

51. Le taux de non-réponse dans les enquêtes varie beaucoup d'un pays et d'une enquête à l'autre. Dans l'exercice de calcul ci-après, nous avons prévu un taux de non-réponse de 10 %. Il va de soi que les pays devront utiliser le chiffre correspondant le plus près aux taux enregistrés lors des dernières enquêtes nationales.

3.3.7. Taille des échantillons-maîtres

52. Les échantillons-maîtres sont étudiés en détail au chapitre 4. Dans la présente section, l'accent sera mis sur la dimension de l'échantillon à utiliser pour l'échantillon-maître. En bref, un échantillon-maître est un échantillon nombreux d'UPE établi pour les pays qui ont mis en œuvre des programmes continus et intégrés d'enquêtes de grande envergure. Cet échantillon de grande taille a pour but d'offrir un nombre suffisant d'échantillons pour pouvoir réaliser sur plusieurs années de multiples enquêtes sans devoir interroger chaque fois les mêmes déclarants.

53. Comme l'échantillon-maître est conçu de manière à être appliqué dans le cadre de nombreuses enquêtes et par conséquent à englober un grand nombre de sujets, il y a manifestement de nombreuses populations cibles et estimations clés à envisager. À cet égard, la plupart des pays définissent la taille de l'échantillon en se fondant sur deux considérations. La première, qui est évidente, est budgétaire. La seconde est la taille prévue des échantillons qui seront utilisés lors des différentes enquêtes menées pendant l'intervalle couvert par l'échantillon-maître, qui atteint fréquemment jusqu'à dix ans entre deux recensements de population. Ainsi, la taille des échantillons-maîtres peut être considérable et atteindre jusqu'à 50 000 ménages, voire davantage. Des plans d'utilisation de l'ensemble de l'échantillon des ménages sont formulés avec soin.

Exemple

Supposons que l'échantillon-maître, dans le pays A, comprend 50 000 ménages. Cet échantillon-maître doit être utilisé pour trois enquêtes qui ont déjà été prévues et peut-être aussi pour deux autres qui ne le sont pas encore. L'une des enquêtes doit porter sur les revenus et les dépenses des ménages et doit être répétée trois fois pendant les dix années suivantes, les première, cinquième et huitième années. Cette enquête doit englober 8 000 ménages à chaque occasion. À la cinquième année, cependant, il y aura un échantillon de remplacement de 4 000 ménages pour la moitié des 8 000 ménages interrogés la première année. De même, la huitième année, les 4 000 ménages restants de la première année seront remplacés par 4 000 autres. Ainsi, l'on interrogera en tout 16 000 ménages pour l'enquête sur les recettes et les dépenses. La deuxième enquête prévue concerne la santé et devrait englober quelque 10 000 ménages, et la troisième, concernant la participation à la population active, portera sur quelque 12 000 ménages. En

tout, 38 000 ménages sont réservés pour ces trois enquêtes. Il reste par conséquent 12 000 ménages qui peuvent être interrogés pour d'autres enquêtes si besoin est.

3.3.8. Estimation du changement du niveau de référence

54. Dans le cas d'enquêtes qui sont répétées périodiquement, les principaux objectifs des mesures consistent à estimer le changement intervenu d'une enquête à l'autre. En termes statistiques, l'estimation obtenue à la première occasion est le *niveau de référence* d'un indicateur donné, et la différence entre ce niveau et le niveau estimé à la deuxième occasion est le *changement* estimatif. Lorsque l'on estime un changement, il faut généralement, pour pouvoir tirer des conclusions fiables, prendre un échantillon de bien plus grande taille que celui qui est nécessaire pour procéder simplement à une estimation. Tel est le cas en particulier lorsque l'on cherche à mesurer des changements de modeste envergure. Cette méthode d'échantillonnage permet néanmoins de réduire la taille de l'échantillon (et par conséquent le coût) nécessaire pour estimer un changement (voir la section 3.9.2).

3.3.9. Budget de l'enquête

55. Il va peut-être sans dire que le budget de l'enquête ne peut pas être ignoré pour déterminer la taille de l'échantillon à utiliser pour une enquête sur les ménages. Le budget, même s'il n'est pas un paramètre figurant dans le calcul mathématique de la taille de l'échantillon, revêt néanmoins une importance considérable sur le plan pratique.

56. C'est le statisticien qui, compte tenu de chacun des paramètres discutés dans le présent chapitre, procède aux premiers calculs de la taille de l'échantillon. Souvent, cependant, l'échantillon ainsi calculé est trop nombreux eu égard au budget disponible. Lorsque tel est le cas, les concepteurs doivent soit mobiliser des fonds supplémentaires, soit modifier les objectifs de la mesure en réduisant soit la précision requise, soit le nombre de domaines.

57. Il appartient au spécialiste de l'échantillonnage de guider la discussion concernant l'arbitrage à opérer entre coût et précision. Il doit expliquer les compromis que suppose une réduction du nombre de domaines (moins d'utilité pour les usagers) ou une réduction de la précision (moins de fiabilité pour les indicateurs clés) lorsque la taille de l'échantillon doit être réduite en raison de considérations budgétaires. La discussion doit aller dans le sens des exemples donnés ci-dessus à propos de la précision et des domaines. Le fait que le nombre de grappes constitue également l'un des principaux éléments déterminants des coûts et de la précision de l'enquête doit aussi être dûment pris en considération (cette question est examinée de manière plus approfondie dans la section 3.5.5).

3.3.10. Calcul de la taille de l'échantillon

58. Nous allons, dans la présente section, indiquer la formule à utiliser pour calculer la taille de l'échantillon compte tenu des paramètres discutés précédemment. Comme il s'agit d'enquêtes sur les ménages, la taille de l'échantillon est le nombre de ménages qui doivent être sélectionnés. Il est également donné des exemples.

59. D'une manière générale, lorsqu'il est inclus une proportion , la formule d'estimation[5] de la taille de l'échantillon, n_h, est :

$$n_h = (z^2)\ (r)\ (1-r)\ (f)\ (k)/(p)\ (\bar{n})\ (e^2) \qquad (3.7)$$

où n_h est le paramètre calculé et la taille de l'échantillon, c'est-à-dire le nombre de ménages à sélectionner; z est la statistique qui définit le niveau de confiance requis, r est une estimation de l'un des indicateurs clés à mesurer lors de l'enquête; f est l'effet imputable à la conception de l'échantillon (*deff*), supposé comme étant de 2,0 (valeur par défaut); k est le multiplicateur visant à tenir compte du taux prévu de non-réponse; p est la proportion de la population totale représentée par la population cible sur laquelle est fondé le paramètre r; $\bar{n}$ est la taille moyenne (nombre de personnes par ménage); et e la marge d'erreur à ne pas dépasser. Les valeurs recommandées de certains des paramètres sont les suivantes.

60. La statistique z à utiliser devrait être de 1,96 pour un degré de confiance de 95 % (elle serait de, par exemple, 1,645 pour un degré de confiance de 90 %). La première valeur est généralement considérée comme la norme pour déterminer le niveau de confiance à atteindre pour évaluer la marge d'erreur dans une enquête sur les ménages. La valeur par défaut de l'effet imputable à la conception de l'échantillon est habituellement considérée comme étant de 2,0, à moins que les données empiriques provenant d'enquêtes précédentes ou d'enquêtes semblables ne conduisent à retenir une autre valeur. Le multiplicateur de non-réponse, k, doit être choisi à la lumière de l'expérience acquise à cet égard et il est habituellement inférieur à 10 % dans les pays en développement. Une valeur de 1,1 pour k serait par conséquent un choix prudent. Le paramètre p peut habituellement être calculé sur la base des résultats du dernier recensement. Le paramètre $\bar{n}$ est souvent de 6,0 dans la plupart des pays en développement mais sa valeur exacte, qui peut habituellement être tirée du dernier recensement, est celle qu'il conviendra d'utiliser. Pour la marge d'erreur, e, il est recommandé de fixer le niveau de précision à 10 % de r; par conséquent, $e = 0{,}10r$. L'on peut prendre un échantillon de taille plus réduite si la marge d'erreur est moins rigoureuse, par exemple, $e = 0{,}15r$, mais les résultats de l'enquête seront évidemment beaucoup moins fiables. En remplaçant certaines valeurs sélectionnées, l'on obtient :

$$n_h = (3{,}84)\ (1 - r)\ (1{,}2)\ (1{,}1)/(r)\ (p)\ (6)\ (0{,}01) \qquad (3.8)$$

L'équation (3.8) se ramène à :

$$n_h = (84{,}5)\ (1 - r)/(r)\ (p) \qquad (3.9)$$

61. La version réduite peut être utilisée lorsque *toutes* les valeurs par défaut recommandées pour les paramètres sont utilisées plutôt que les valeurs plus précises qui peuvent être tirées de l'expérience nationale.

5 La formule peut également comporter un facteur, appelé multiplicateur infini, dont il doit être tenu compte lorsqu'il s'avère que la taille calculée de l'échantillon représente une proportion significative de la population. Cependant, tel est rarement le cas des enquêtes de grande envergure sur les ménages du type envisagé dans le présent guide. En conséquence, le multiplicateur infini est supposé avoir une valeur égale à l'unité et est par conséquent ignoré dans la formule 3.5.

Exemple

Dans le pays B, il est décidé que le principal indicateur à mesurer est le taux de chômage, que l'on pense être d'environ 10 % de la population active civile. La population active civile est définie comme étant la population de 14 ans et plus, représentant 65 % environ de la population totale du pays. En l'occurrence, r = 0,1 et p = 0,65. Supposons que nous souhaitions estimer le taux de chômage avec une marge d'erreur relative de 10 % au niveau de confiance de 95 %; alors, $e = 0{,}10r$ (c'est-à-dire une erreur type de 0,01), comme recommandé ci-dessus. En outre, les valeurs du taux prévu de non-réponse, de l'effet dû à la conception et de la taille moyenne des ménages sont celles que nous avons recommandées. Nous pouvons alors utiliser la formule (3.9), qui donne 1 170 ménages [(84,5*0,9)/(0,1*0,65)]. C'est là un échantillon de taille assez réduite, essentiellement parce que la population de base constitue une proportion si grande du total, c'est-à-dire 65 %. Il y a lieu de rappeler que la dimension de l'échantillon ainsi calculée vaut pour un seul domaine, en l'occurrence le domaine national. Si les objectifs de la mesure sont aussi d'obtenir des données également fiables pour les régions urbaines et rurales, il faudra doubler la taille de l'échantillon, à supposer que tous les paramètres des formules (3.8) et (3.9) s'appliquent aux milieux aussi bien urbain que rural. Si ces paramètres diffèrent (par exemple si la taille moyenne des ménages urbains n'est pas identique à celle des ménages ruraux, ou si les taux de non-réponse sont différents en milieu urbain et en milieu rural), il faudra utiliser les valeurs les plus exactes pour calculer séparément les dimensions des échantillons à utiliser en milieu urbain et en milieu rural. Les résultats seraient évidemment différents.

62. L'exemple suivant envisage une population de base plus restreinte, à savoir les enfants de moins de 5 ans.

Exemple

Dans le pays C, le principal indicateur à mesurer est le taux de mortalité des enfants de moins de 5 ans, que l'on pense être d'environ 5 %. En l'occurrence, r = 0,05 et p est estimé comme étant d'environ 0,15, soit 0,03*5. Dans ce cas également, nous souhaitons estimer le taux de mortalité avec une marge d'erreur relative de 10 %; alors, $e = 0{,}10r$ (soit une erreur type de 0,005). Les valeurs pour le taux prévu de non-réponse, l'effet imputable à la conception de l'échantillon et la taille moyenne des ménages sont à nouveau celles que nous recommandées. La formule (3.9) donne près de 10 704 ménages (84,5*0,95)/(0,05*0,15), soit un échantillon de taille beaucoup plus importante que celui de l'exemple précédent. Dans ce cas également, la principale raison tient à la taille de la population de base, c'est-à-dire les enfants de moins de 5 ans, qui ne constituent que 15 % du total. Le paramètre r est réduit aussi, ce qui, avec un p réduit, débouche inévitablement sur un échantillon de grande taille.

63. Le dernier exemple envisage le cas où la population cible est la population totale. Ici, p = 1 et peut être ignoré; cependant, les formules (3.8) et (3.9) peuvent continuer d'être utilisées si l'on retient les valeurs recommandées pour les paramètres.

Exemple

Dans le pays D, le principal indicateur à évaluer est la proportion par rapport à la population totale que représentent les personnes qui ont eu un sérieux problème de santé au cours de la semaine écoulée. Cette proportion, pense-t-on, est comprise entre 5 et 10 %, et il est par conséquent utilisé le plus faible de ce chiffre, étant donné qu'il donnera un échantillon de plus grande dimension (l'approche prudente). En l'occurrence, r = 0,05 et p est évidemment égal

à 1,0. Une fois de plus, nous souhaitons estimer l'incidence des graves problèmes de santé avec une marge d'erreur relative de 10 % : $e = 0{,}10r$ [6] (soit une erreur type de 0,005), et les valeurs du taux escompté de non-réponse, de l'effet imputable à la conception de l'échantillon et la dimension moyenne des ménages sont celles que nous avons recommandées. La formule (3.9) donne un peu plus de 1 600 ménages (84,5*0,95)/(0,05).

64. Comme on l'a dit, la dimension de l'échantillon utilisé pour l'enquête peut en définitive être déterminée en calculant les dimensions des échantillons à utiliser pour plusieurs indicateurs clés et en retenant celui qui donne l'échantillon le plus nombreux. En outre, avant de parvenir à une décision finale, il faudra prendre en compte aussi le nombre de domaines visés par l'enquête ainsi que le budget disponible.

65. Pour les pays dans le cas desquels une ou plusieurs des hypothèses susmentionnées ne sont pas valides, l'on peut, dans la formule (3.7), procéder à des substitutions simples pour obtenir des chiffres plus exacts concernant la dimension de l'échantillon. Il se peut par exemple que la taille moyenne des ménages soit inférieure ou supérieure à 6,0, que le taux de non-réponse à prévoir soit environ de 5 % plutôt que de 10 % et que la valeur de p pour le pays puisse être calculée de façon plus précise au moyen des chiffres provenant du recensement.

66. Toutefois, il est recommandé de ne pas modifier la valeur statistique z de 1,96, qui est la norme communément utilisée. Il convient également de conserver pour l'effet dû à la conception de l'échantillon, f, une valeur de 2,0, à moins que, comme on l'a vu, des données récentes provenant d'une autre source conduisent à choisir une autre valeur. Il est également recommandé que e soit défini comme étant égal à $0{,}10r$, sauf si, pour des raisons budgétaires, il n'est pas possible de retenir un échantillon de cette taille, auquel cas cette valeur pourrait être portée à $0{,}12r$ ou à $0{,}15r$. Ces augmentations de la marge d'erreur, cependant, se traduiront par des erreurs d'échantillonnage beaucoup plus élevées.

3.4. Stratification

67. La stratification de la population à interroger avant de sélectionner l'échantillon est une méthode communément utilisée lors de la conception d'une enquête sur les ménages. La stratification permet de classer la population en sous-populations (strates) sur la base d'informations auxiliaires disponibles au sujet de l'ensemble de la population. Les éléments de l'échantillon sont alors sélectionnés indépendamment, dans chaque strate, à la lumière des caractéristiques à mesurer.

3.4.1. Stratification et allocation de l'échantillon

68. Avec un échantillonnage stratifié, les dimensions de l'échantillon à l'intérieur de chaque strate dépendent du technicien plutôt que d'un choix aléatoire fondé sur le processus d'échantillonnage. Une population divisée en strates peut avoir exactement n_s unités d'échantillonnage dans chaque strate, où n_s est le nombre souhaité d'unités d'échantillonnage dans la $s^{ième}$ strate. Un échantillon non stratifié, en revanche, donnerait pour la $s^{ième}$ sous-population un échantillon de taille un peu différente de n_s.

6 Comme r s'applique en l'occurrence à l'ensemble de la population, il est égal à p, de sorte que e est égal à $0{,}10p$.

Exemple

Supposons que l'échantillon retenu pour une enquête se compose de deux strates, une strate urbaine et une strate rurale. Des informations provenant du recensement de la population sont disponibles pour classer toutes les circonscriptions administratives en circonscriptions urbaines ou rurales, ce qui permet de stratifier la population selon ce critère. Il est décidé de sélectionner un échantillon proportionné (par opposition à disproportionné) dans chaque strate étant donné que la population est composée à concurrence de 60 % de populations rurales et de 40 % de populations urbaines. Si la taille de l'échantillon est de 5 000 ménages, une sélection indépendante de l'échantillon par strates donnera 3 000 ménages ruraux et 2 000 ménages urbains. Si l'échantillon a été sélectionné de façon aléatoire sans stratification préalable, la répartition des ménages faisant partie de l'échantillon ne sera pas 3 000-2 000, bien que cela soit la répartition prévisible. L'échantillon non stratifié pourrait, par malchance, donner un échantillon composé, par exemple, de 3 200 ménages ruraux et de 2 800 ménages urbains.

69. L'une des justifications de la stratification est donc de réduire ce risque de malchance et d'avoir un nombre excessivement important (ou réduit) d'unités sélectionnées parmi une sous-population considérée comme importante aux fins de l'analyse. La stratification a pour but de garantir une représentation appropriée des importants groupes de sous-populations sans introduire de distorsions dans l'opération de sélection. Il importe de noter toutefois que représentation appropriée n'est pas synonyme d'échantillonnage proportionné. Fréquemment, une ou plusieurs des strates peuvent également être des demandes d'estimation (voir ci-dessus), auquel cas il peut être nécessaire de choisir des échantillons de tailles égales dans la strate considérée, ce qui donnera un échantillon disproportionné par strate. Une *allocation* à la fois proportionnée et disproportionnée des unités d'échantillonnage aux différentes strates constitue par conséquent un aspect légitime de la conception d'un échantillon stratifié, et le choix dépendra des objectifs de la mesure.

70. Comme le donne à entendre la phrase précédente, la stratification peut également être un moyen d'allouer l'échantillon implicitement, méthode plus simple et plus pratique qu'une allocation optimale[7]. Autrement dit, avec un échantillonnage proportionné par strates, il n'est pas nécessaire de calculer à l'avance le nombre d'échantillons à allouer à chaque strate.

Exemple

Supposons que l'échantillon soit conçu de manière à assurer de façon précise une allocation proportionnée de l'échantillon total de chacune des dix provinces qui constituent le pays. Si, par exemple, la province A représente 12 % de la population totale, il faudra sélectionner dans ces provinces 12 % des grappes d'échantillons, à condition que la taille des grappes soit constante. Supposons en outre que le nombre total de grappes à sélectionner dans l'ensemble du pays soit de 400. Une méthode fréquemment utilisée dans de nombreux pays consiste à *affecter* 48 grappes (0,12 * 400) à la province A. Avec une stratification appropriée, cependant, cette procédure est superflue, chaque province devant plutôt être considérée comme une strate séparée dans le cadre du processus de sélection de l'échantillon. Ainsi, l'application d'un échan-

7 Par allocation optimale, l'on entend une allocation fondée sur les fonctions de coût et les différentes variances à l'intérieur de chaque strate (mesures d'hétérogénéité). Cette question n'est pas abordée dans le présent guide car cette méthode est rarement utilisée dans la pratique dans les pays en développement, peut-être parce que l'on ne dispose pas d'estimations finales des coûts des enquêtes. Le lecteur trouvera des informations détaillées concernant l'allocation optimale dans nombre des ouvrages de référence énumérés à la fin de ce chapitre.

tillonnage systématique avec probabilité proportionnelle à la taille (voir le tableau 3.1), avec un seul intervalle d'échantillonnage, donnera automatiquement 48 grappes dans la province A. Ce type de stratification ainsi que son utilisation pour simplifier les systèmes d'allocation sont examinés de manière plus approfondie dans la section 3.4.3.

3.4.2. Règles de stratification

71. Pour stratifier une population, il y a deux règles fondamentales à suivre. L'une d'elles doit toujours être respectée. L'autre doit normalement l'être aussi, bien que son inobservation n'affecte guère la conception de l'échantillon. La règle impérative est qu'au moins une unité d'échantillonnage soit sélectionnée parmi chacune des strates créées. Les strates sont essentiellement des sous-groupes indépendants et s'excluant mutuellement de la population : chaque élément d'une population peut appartenir à une strate et à une seule. Du fait de cette caractéristique, des échantillons doivent obligatoirement être prélevés dans chaque strate de manière à pouvoir sonder l'ensemble de la population et parvenir à une estimation de la moyenne de la population ne présentant pas de distorsion. Comme, théoriquement, chaque strate peut être traitée indépendamment lors de la conception de l'échantillon, il n'est pas nécessaire que les strates soient créées sur la base de critères objectifs. Des critères subjectifs, si on le souhaite, peuvent être utilisés aussi. Le principe directeur doit être que les unités constituant une strate doivent être aussi semblables que possibles au regard de la variable A étudiée de manière à réduire la variabilité à l'intérieur de chaque strate.

72. La deuxième règle est que chacune des strates créées doit, idéalement, être aussi différente que possible des autres. L'hétérogénéité *entre* strates et l'homogénéité *à l'intérieur* des strates doivent par conséquent être les principales caractéristiques à rechercher dans l'établissement de strates. L'on voit donc aisément pourquoi les régions urbaines et rurales sont fréquemment considérées comme deux des strates utilisées pour une enquête sur les ménages. Comme indiqué ci-dessus, les populations urbaines et rurales diffèrent à bien des égards (type d'emploi, source et montant des revenus, taille moyenne des ménages, taux de fécondité, etc.), tandis que les membres de la population constituant chacun de ces sous-groupes se ressemblent.

73. L'aspect hétérogénéité est un guide utile pour déterminer combien de strates il y a lieu de créer. Il ne devrait pas y avoir plus de strates qu'il n'y a de sous-populations identifiables dans le contexte du critère spécifique utilisé pour les définir. Par exemple, si un pays est subdivisé en huit régions administratives distinctes et si deux des régions sont très semblables dans le contexte de l'objet de l'enquête, l'échantillon pourrait être conçu sur la base de sept strates (les deux régions similaires étant combinées). Il n'y a aucun intérêt à créer, par exemple, 20 strates si 10 peuvent donner les mêmes sous-groupes hétérogènes.

74. Il importe de noter qu'en ce qui concerne la sélection *proportionnée,* l'échantillon obtenu est au moins aussi précis qu'un échantillon aléatoire simple de même taille. Ainsi, la stratification améliore la précision ou la fiabilité des estimations, et les différences sont les plus marquées lorsque l'hétérogénéité entre les strates est la plus grande. C'est cet aspect de l'échantillonnage stratifié qui fait que même une mauvaise stratification[8] ne compromet pas la fiabilité des estimations.

8 La stratification peut être mauvaise lorsque des strates sont créées inutilement ou lorsque certains éléments de la population sont classés dans les strates erronées.

75. Un autre point important a trait à l'estimation des erreurs d'échantillonnage. Si une seule unité sélectionnée dans chaque strate suffit pour que les conditions inhérentes à un échantillonnage stratifié soient théoriquement réunies, il faut en choisir au *minimum deux* si l'on veut utiliser les résultats pour calculer les erreurs d'échantillonnage qui caractérisent les estimations.

76. Il peut parfois être nécessaire d'utiliser de nombreuses variables aux fins de la stratification. En pareils cas, il faut s'en tenir aux principes suivants : il est préférable que les variables de stratification soient indépendantes les unes des autres mais liées à la variable visée par l'enquête; les cellules constituées n'ont pas à être complètes (les cellules les plus petites et les moins importantes peuvent être combinées); et, d'une manière générale, il est préférable d'utiliser des groupements plus approximatifs de nombreuses variables que des groupements plus précis d'une seule variable.

3.4.3. Stratification implicite

77. Comme indiqué ci-dessus, le choix des informations disponibles pour créer des strates est dicté par les objectifs de mesure de l'enquête. Pour les enquêtes de grande envergure sur les ménages et de caractère général, une méthode particulièrement utile est celle dite de la stratification *implicite.* Le fait que son critère essentiel est géographique suffit généralement à répartir convenablement l'échantillon parmi les sous-groupes les plus importants de la population, comme la population urbaine et rurale, les régions administratives, les sous-populations ethniques, les groupes socioéconomiques, etc. Du fait de cette caractéristique géographique, la stratification implicite est extrêmement utile aussi lorsque l'objet de l'enquête est un thème unique, que ce soit la population active, l'activité économique des ménages, la mesure de la pauvreté, la santé ou les recettes et les dépenses. Cette technique est vivement recommandée pour ces raisons et aussi en raison de sa simplicité d'implication.

78. Pour être appliquée correctement, la stratification implicite exige une sélection systématique au niveau primaire. La procédure est simple et consiste à commencer par organiser le fichier d'UPE dans l'ordre géographique. Dans de nombreux pays, cet ordre est généralement urbain, par province, puis, à l'intérieur de chaque province, par district, et ensuite rural par province et, à l'intérieur de chaque province, par district. L'étape suivante consiste à sélectionner systématiquement l'UPE dans le fichier ainsi trié. La sélection systématique est effectuée soit par échantillonnage à probabilité égale soit, ce qui est plus fréquent, par échantillonnage avec probabilité proportionnelle à la taille.

79. Comme on l'a déjà dit, un avantage important de la stratification implicite est qu'elle élimine la nécessité de créer des strates géographiques explicites, ce qui, à son tour, élimine celle d'allouer l'échantillon à ces strates, surtout lorsqu'il est utilisé un échantillonnage proportionné. Un autre avantage est la simplicité décrite dans le paragraphe précédent étant donné que cette méthode exige simplement un tri du fichier et l'application de l'intervalle ou des intervalles des échantillonnages. Un échantillonnage disproportionné peut aussi être appliqué aisément au niveau primaire du tri géographique. Par exemple, si le premier niveau est constitué par les secteurs urbain et rural, il suffit d'appliquer des taux d'échantillonnage différents aux segments urbain et rural. La figure 3.1 illustre un système de stratification implicite avec échantillonnage systématique. L'échantillonnage avec probabilité proportionnelle à la taille est discuté plus en détail dans la section 3.7 ci-après.

Figure 3.1
Organisation des circonscriptions administratives en vue d'une stratification implicite

Urbain			
	Province 01		
		District 01	
			ZE 001
			ZE 002
			ZE 003
			ZE 004
		District 02	
			ZE 005
			ZE 006
			ZE 007
	Province 02		
		District 01	
			ZE 008
			ZE 009
		District 02	
			ZE 010
			ZE 011
			ZE 012
	Province 03, etc.		
Rural			
	Province 01		
		District 01	
			ZE 101
			ZE 102
			ZE 103
			ZE 104
		District 02	
			ZE 105
			ZE 106
			ZE 107
	Province 02		
		District 01	
			ZE 108
			ZE 109
			ZE 110
			ZE 111
		District 02	
			ZE 112
			ZE 113
			ZE 114
	Province 03, etc.		

3.5. Échantillonnage en grappes

80. L'expression « échantillonnage en grappes » a initialement été conçue pour désigner les échantillons conçus de manière que tous les membres d'un groupe soient interrogés, les groupes eux-mêmes étant définis comme étant les grappes. Par exemple, il est sélectionné un échantillon d'écoles au niveau primaire et de classes au niveau secondaire. Si tous les élèves de chaque classe sont interrogés, nous aurons une grappe de classes. Dans les enquêtes sur les ménages, un exemple d'échantillonnage en grappes du type initial serait la sélection de pâtés de maisons dont tous les habitants seraient interrogés. Ces dernières années, cependant, l'expression « échantillonnage en grappes » a été utilisée de manière plus large pour désigner, d'une façon plus générale, les enquêtes comportant une avant-dernière étape d'échantillonnage qui sélectionne (et définit) les grappes, comme villages, zones d'énumération du recensement ou pâtés de maisons. Lors de la dernière étape de l'échantillonnage, un sous-échantillon de ménages de chaque grappe sélectionnée sont interrogés, et non pas tous. Cette dernière utilisation de cette expression est celle qui est généralement employée dans ce guide.

81. Dans les enquêtes sur les ménages, la conception de l'échantillon repose invariablement et inévitablement, sous une forme ou sous une autre, sur un échantillonnage en grappes, le but étant de contenir les coûts. Comme indiqué plus haut, il est beaucoup moins cher de mener une enquête auprès de 1 000 ménages de 50 localités (20 ménages par grappe) qu'auprès de 1 000 ménages sélectionnés au hasard parmi l'ensemble de la population. Regrettablement, la mise en grappes de l'échantillon affecte sa fiabilité étant donné que les personnes qui constituent la même grappe tendent à être homogènes ou à présenter des caractéristiques plus ou moins semblables. Cet effet dit de mise en grappes doit être compensé, lors de la conception de l'échantillon, par une augmentation correspondante de la taille de l'échantillon.

3.5.1. Caractéristiques de l'échantillonnage en grappes

82. L'échantillonnage en grappes s'écarte nettement de l'échantillonnage stratifié à deux égards[9]. Dans ce dernier cas, toutes les strates sont représentées dans l'échantillon étant donné qu'il est sélectionné un échantillon d'unités dans chaque strate. Dans l'échantillonnage en grappes, il est procédé à une sélection des grappes elles-mêmes, de sorte que celles qui font partie de l'échantillon représentent celles qui en sont exclues. Cette première différence distinctive entre l'échantillonnage stratifié et l'échantillonnage en grappes amène à la deuxième différence. Comme on l'a vu, les strates doivent idéalement être créées de manière à être homogènes au plan interne et hétérogènes au plan externe eu égard aux variables à mesurer. C'est l'inverse qui est vrai dans le cas des grappes. Pour améliorer la précision de l'échantillon, il est préférable d'avoir des grappes qui soient aussi hétérogènes au plan interne que possible.

83. Le fait que, dans les enquêtes sur les ménages, les grappes sont presque toujours définies comme des unités géographiques comme villages ou quartiers de villages signifie regrettablement qu'il n'est généralement pas possible de parvenir à un degré élevé d'hétérogénéité à l'intérieur de la grappe. En fait, des grappes géographiquement définies sont souvent plus homogènes qu'hétérogènes au plan interne pour ce qui est de variables comme le type d'emploi (agriculture, par exemple), le niveau de

9 Il importe de noter que l'échantillonnage stratifié et l'échantillonnage en grappes ne sont pas des méthodes concurrentes étant donné que l'une et l'autre sont invariablement utilisées pour l'échantillonnage lors d'enquêtes sur les ménages.

revenu, etc. La mesure dans laquelle les grappes sont homogènes pour une variable donnée détermine par conséquent l'étendue de la « mise en grappes ». Plus l'échantillon est mis en grappes et moindre est sa fiabilité.

3.5.2. Effet de mise en grappes

84. L'effet de mise en grappes d'un échantillon est mesuré en partie par l'effet de conception (*deff*). Cependant, le *deff* reflète également les effets de la stratification. L'équipe chargée de concevoir l'échantillon doit par conséquent veiller à ce que celui-ci soit conçu de manière à concilier au mieux la nécessité de minimiser les coûts et celle de maximiser la précision. À cette fin, il faut minimiser ou maîtriser autant que possible l'effet de conception. Pour déterminer comment la composante mise en grappes du *deff* peut être minimisée ou maîtrisée, il est bon de se référer à sa définition mathématique :

$$deff = 1 + \delta(\tilde{n} - 1), \tag{3.10}$$

où δ est la corrélation intraclasse (ou intra-grappe), c'est-à-dire la mesure dans laquelle deux unités d'une grappe, en comparaison de deux unités sélectionnées au hasard parmi la population, risquent d'avoir la même valeur, et $\tilde{n}$ est le nombre d'unités de la population cible que comporte la grappe.

85. L'équation (3.10) n'est pas, à strictement parler, la formule du *deff* car elle méconnaît la stratification ainsi qu'un autre facteur qui est introduit lorsque la taille des grappes n'est pas uniforme. Néanmoins, comme la composante mise en grappes est le facteur déterminant dans le *deff,* il peut être utilisé, de façon approximative, pour montrer comment la mise en grappes affecte la conception de l'échantillon et ce qui peut être fait pour y remédier.

86. Il découle de l'équation ci-dessus que le *deff* est un multiplicateur de deux variables, la corrélation intraclasse, δ, et la taille de l'échantillon, $\tilde{n}$. Ainsi, le *deff* augmente parallèlement à l'accroissement aussi bien de δ que de $\tilde{n}$. Si l'enquêteur n'a aucun contrôle sur la corrélation intraclasse de la variable à l'examen, il peut modifier la taille de la grappe, en l'augmentant ou en la diminuant, et ainsi, pour une large part, maîtriser l'effet de conception.

Exemple

Supposons qu'une population ait une corrélation intraclasse de 0,03, c'est-à-dire une corrélation relativement réduite, en ce qui concerne les maladies chroniques. Supposons en outre que les concepteurs de l'échantillon cherchent à déterminer s'il y a lieu d'utiliser des grappes de 10 ou de 20 ménages, avec un échantillon global de 5 000 ménages. Supposons par ailleurs, pour simplifier l'exemple, que tous les ménages aient la même taille et comptent 5 personnes. La valeur de $\tilde{n}$ est alors de 50 pour 10 ménages et de 100 pour 20 ménages. Une substitution simple dans l'équation (3.10) donne pour le *deff* une valeur approximative de [1 + 0,03(49)], ou 2,5 pour la grappe de 10 ménages mais de 4,0 pour la grappe de 20 ménages. Ainsi, l'effet de conception augmente en gros de 60 % dans le cas de la grappe de plus grande taille. Les concepteurs devront alors déterminer laquelle des deux options est préférable : il faut interroger deux fois plus de grappes (500) en utilisant l'option de 10 ménages pour maintenir la fiabilité dans des limites acceptables ou choisir l'option, moins onéreuse, de 250 ménages, au prix d'accroître considérablement la variance de l'échantillonnage. Il va de soi que l'on peut également envisager d'autres options comprises entre 10 et 20 ménages.

87. Il y a plusieurs façons d'interpréter l'effet de conception : comme étant le facteur égal à la différence entre la variance de l'échantillon devant effectivement être utilisé pour l'enquête par rapport à celle d'un échantillon aléatoire simple de même taille; simplement comme une mesure de la moindre précision de l'échantillon effectif par rapport à l'échantillon aléatoire simple; ou comme un reflet du nombre de grappes supplémentaires à sélectionner en comparaison d'un échantillon aléatoire simple pour obtenir une variance équivalente. Par exemple, un *deff* de 2,0 signifie qu'il faut sélectionner deux fois plus de cas pour obtenir une fiabilité semblable à celle que donnerait un échantillon aléatoire simple. Il est donc manifestement déconseillé de concevoir un plan d'échantillonnage dont le *deff* dépasserait de beaucoup 2,5 à 3,0 pour les indicateurs clés.

3.5.3. Taille des grappes

88. Comme on l'a vu, le concepteur n'a pas de contrôle sur les corrélations. De plus, pour la plupart des variables, il n'existe guère de recherche empirique qui ait essayé d'estimer la valeur de ces corrélations. La corrélation intraclasse peut varier, théoriquement, entre – 1 et + 1, bien qu'il soit difficile de concevoir beaucoup de variables pour lesquelles elle serait négative. La seule possibilité pour maintenir le *deff* au minimum consiste par conséquent à faire en sorte que la taille des grappes soit aussi réduite que le permet le budget. Le tableau 3.3 indique les *deffs* pour différentes valeurs de la corrélation intraclasse lorsque la taille de la grappe est constante.

Tableau 3.3
Comparaison des composantes mise en grappes de l'effet de conception pour différentes corrélations intraclasses δ et tailles de grappes $\tilde{n}$

	δ						
	0,02	0,05	0,10	0,15	0,20	0,35	0,50
5	1,08	1,20	1,40	1,60	1,80	2,40	3,00
10	1,18	1,45	1,90	2,35	2,80	4,15	5,50
20	1,38	1,95	2,90	3,85	4,80	6,65	10,50
30	1,58	2,45	3,90	5,35	6,80	11,15	15,50
50	1,98	3,45	5,90	8,35	10,80	18,15	25,50
75	2,48	4,70	8,40	12,10	15,80	26,90	38,00

89. Le tableau 3.3 montre clairement que des grappes d'une taille supérieure à 20 donneront des *deff*s inacceptables (supérieurs à 3,0), à moins que la corrélation intraclasse soit très réduite. Pour évaluer les chiffres du tableau, il importe de ne pas perdre de vue que $\tilde{n}$ désigne le nombre d'unités de la population cible et non le nombre de ménages. À cet égard, la valeur de $\tilde{n}$ à utiliser est égale au nombre de ménages que comporte la grappe, multiplié par le nombre moyen de personnes que comptent les groupes cibles. Si le groupe cible, par exemple, est les femmes de 14 à 49 ans, il y a habituellement pour ce groupe une femme par ménage, auquel cas une grappe de b ménages aura approximativement ce même nombre de femmes de 14 à 49 ans. Autrement dit, $\tilde{n}$ et b sont à peu près égaux pour ce groupe cible, et le tableau 3.3 s'applique tel quel. Dans l'exemple qui suit, le nombre de ménages et la population cible ne sont pas égaux.

Exemple

Supposons que la population cible soit toutes les personnes, comme ce serait le cas d'une enquête sur la santé ayant pour but d'estimer l'incidences des maladies aiguës et chroniques.

Supposons en outre que l'enquête doive utiliser des grappes de 10 ménages. En l'occurrence, la valeur de $\bar{n}$ est égale à dix fois la taille du ménage moyen; si celle-ci est de 5,0, $\bar{n}$ est égal à 50. Ainsi, 50 est la valeur de $\bar{n}$ qui doit être utilisée dans le tableau 3.3 pour évaluer son *deff* potentiel. Le tableau 3.3 fait apparaître que le *deff* est très important sauf lorsque δ est d'environ 0,02. Cela porte à penser qu'un échantillon en grappes conçu de manière à n'utiliser que 10 ménages par grappe donnerait des résultats très peu fiables pour une caractéristique comme les maladies contagieuses étant donné que cette caractéristique aurait un δ important.

90. L'exemple illustre pourquoi il est si important de tenir compte de la taille de la grappe lors de la conception d'une enquête sur les ménages, particulièrement pour les indicateurs clés à mesurer. De plus, il ne faut pas perdre de vue que la taille de la grappe, dans la description de la conception de l'échantillon, se rapportera généralement au nombre de ménages, tandis que la taille de la grappe, aux fins de l'évaluation et des effets de conception, sera plutôt liée à la population ou aux populations cibles.

3.5.4. Calcul de l'effet de conception (*deff*)

91. La valeur du *deff* pour les variables spécifiées par les analystes peut être calculée après la fin de l'enquête. Il faut pour cela estimer la variance de l'échantillonnage à partir des variables sélectionnées (les méthodes sont discutées au chapitre 7) puis calculer pour chaque variable le ratio entre sa variance et celle d'un échantillon aléatoire simple de même taille. Ce calcul est une estimation du *deff* « complet », y compris les effets de stratification ainsi que la variabilité due à la taille des grappes, et pas seulement l'élément mise en grappes.

92. La racine carrée du ratio des variances donne le ratio des erreurs types, c'est-à-dire le *deff,* ce qui est souvent calculé dans la pratique et présenté dans la documentation technique d'enquêtes comme les enquêtes démographiques et sanitaires (DHS).

3.5.5. Nombre de grappes

93. Il importe de se rappeler que la taille de la grappe est importante, non seulement en raison de son impact sur la précision de l'échantillonnage, mais aussi dans le contexte de la taille de l'échantillon global car elle détermine le nombre de zones différentes qui doivent être visitées lors de l'enquête. Le fait que cela affecte directement le coût de l'enquête est précisément la raison pour laquelle l'on a recours à des échantillons en grappes. Ainsi, un échantillon de 10 000 ménages comportant des grappes de 10 ménages chacun exigera 1 000 grappes, tandis que des grappes de 20 ménages n'en exigeront que 500. Comme indiqué plus haut, il importe au plus haut point de tenir compte des considérations touchant aussi bien le coût que la précision pour prendre une décision sur cet aspect de la conception de l'échantillon.

3.6. Échantillonnage par phases

94. Du point de vue théorique, le plan idéal consiste à sélectionner l'échantillon de ménages, *n,* au hasard parmi les strates dûment identifiées qui constituent l'ensemble de la population de ménages, *N*. L'échantillon aléatoire stratifié ainsi obtenu donnerait le maximum de précision. Toutefois,

utiliser un échantillon de ce type coûte beaucoup trop cher pour être possible[10], comme nous l'avons vu dans le contexte des économies que permet de réaliser un échantillonnage en grappes.

3.6.1. Avantages d'un échantillonnage par phases

95. La sélection d'un échantillon en plusieurs *phases* a des avantages pratiques pour le processus de sélection lui-même. Elle permet d'isoler, par phases successives, les zones géographiques où sera menée l'enquête et en particulier d'établir une liste des ménages et de mener les entrevues. Lorsqu'il faut établir une liste parce que le cadre d'échantillonnage est obsolète, l'on peut introduire une phase de sélection pour limiter la taille de l'échantillon à interroger.

96. Avec un échantillonnage en grappes, la procédure de sélection comporte généralement au minimum deux phases, premièrement la sélection des grappes et, deuxièmement, la sélection des ménages. Dans les enquêtes sur les ménages, les grappes sont toujours définies comme des unités géographiques d'un type ou d'un autre. Si ces unités sont suffisamment réduites, aussi bien géographiquement que par les effectifs de leur population, et s'il peut être établi une liste à jour, complète et exacte des unités parmi lesquelles il peut être sélectionné un échantillon, ces deux phases peuvent suffire pour établir le plan. Si la plus petite des unités géographiques disponibles est trop grande pour pouvoir être utilisée efficacement comme grappe, il faudra prévoir une sélection en trois phases.

Exemple

Supposons qu'un pays souhaite définir ses grappes comme étant les zones d'énumération (ZE) du recensement, celles-ci étant les circonscriptions géographiques les plus réduites sur le plan administratif. Le *cadre* de ZE (voir le chapitre 4 pour une discussion plus détaillée des cadres) est complet car tout le pays est divisé en ZE. Il est exact parce que, par définition, tous les ménages vivent dans une ZE et une seule. En outre, il est raisonnablement à jour en ce sens qu'il est fondé sur le dernier recensement, à condition que les définitions des ZE n'aient pas changé depuis le dernier recensement. Supposons en outre que celui-ci remonte à deux ans. Il apparaît par conséquent qu'il faudra compiler une liste plus à jour des ménages qui vivent dans les ZE faisant partie de l'échantillon, plutôt que d'utiliser la liste de ménages établie en vue du recensement d'il y a deux ans. La taille moyenne d'une ZE est de 200 ménages, mais il est prévu que les grappes aient une taille de 15 ménages chacune. Les enquêteurs parviennent à la conclusion qu'établir une liste de 200 ménages pour chaque groupe de 15 ménages qui seront en définitive interrogés (soit un ratio de plus de 13 contre 1) coûterait trop cher. Ils décident par conséquent d'utiliser une formule moins onéreuse selon laquelle chaque ZE faisant partie de l'échantillon est divisée en quadrants de taille à peu près égale d'une cinquantaine de ménages chacun. Le plan d'échantillonnage est alors modifié de manière à sélectionner un quadrant ou *segment* dans chaque ZE pour l'établissement de la liste, ce qui réduit la charge de travail des trois quarts. Nous avons ainsi trois phases : première phase, sélection des ZE; deuxième phase, sélection des segments de ZE; et, troisième phase, sélection des ménages.

10 Il y a eu une ou deux exceptions, qui seraient celles de pays géographiquement très exigus, comme le Koweït, où la sélection d'un échantillon aléatoire de ménages n'entraînera que des frais de déplacement très réduits.

3.6.2. Utilisation de phases fictives

97. L'on a fréquemment recours à des phases dites fictives lors de la sélection de l'échantillon, pour éviter d'avoir, lors de l'avant-dernière phase, à sélectionner un échantillon à partir d'un énorme fichier d'unités. Le fichier peut contenir plein d'unités et être si volumineux qu'il ne peut pas, réalistement, être utilisé au moyen d'un fastidieux processus manuel de sélection. Même si le fichier est informatisé, son volume peut être tel qu'il ne peut pas véritablement être utilisé pour la sélection de l'échantillon[11]. Des phases fictives permettent de rétrécir les sous-univers et de les ramener à des dimensions plus gérables en exploitant le caractère hiérarchisé des circonscriptions administratives d'un pays.

98. Dans le cas des enquêtes rurales réalisées au Bangladesh, par exemple, les villages sont souvent sélectionnés à l'avant-dernière phase. Il y a dans le pays plus de 100 000 villages, soit un nombre trop élevé pour pouvoir sélectionner un échantillon. Si le plan d'échantillonnage prévoit de sélectionner 600 villages lors de l'avant-dernière phase, par exemple, il n'en serait sélectionné que 1 sur 167. Pour réduire les dimensions des fichiers devant être utilisés pour la sélection de l'échantillon, il pourra être décidé de sélectionner celui-ci par phases, en se fondant sur la hiérarchie de circonscriptions administratives constituant le Bangladesh : thanas, districts et villages. La sélection d'échantillons se fera par phases : il sera tout d'abord sélectionné 600 thanas, sur la base d'une probabilité proportionnelle à leur taille (cette méthode est discutée en détail dans la section 3.7). Ensuite, il sera sélectionné exactement un district pour chaque thana retenu comme échantillon, dans ce cas également sur la base d'une probabilité proportionnelle à la taille : il y aurait ainsi dans l'échantillon quelque 600 districts. Troisièmement, il serait sélectionné un village sur la base d'une probabilité proportionnelle à la taille, de sorte que l'on aurait à nouveau 600 villages. Enfin, un échantillon de ménages serait sélectionné pour chaque village retenu comme échantillon. Cela donnerait généralement un échantillon systématique de tous les ménages vivant dans chaque village de l'échantillon.

99. La méthode de sélection de l'échantillon décrite ci-dessus constitue en réalité un échantillonnage en deux phases des villages et des ménages, bien qu'il ait d'abord été utilisé deux phases fictives pour sélectionner les thanas et les districts parmi lesquels devaient être sélectionnés les villages. Il faut en pareil cas illustrer mathématiquement le caractère factice des deux premières phases en examinant les probabilités à chaque phase de la sélection ainsi que la probabilité globale.

3.6.2.1. Première phase de la sélection : thanas

100. Les thanas sont sélectionnés sur la base d'une probabilité proportionnelle à la taille. La probabilité à cette phase est donnée par l'équation :

$$P_1 = \left(am_t\right) \Big/ \sum m_t \qquad (3.11)$$

11 L'on peut néanmoins décomposer en sous-fichiers distincts pour chaque strate ou circonscription administrative (une région ou une province) un très volumineux fichier informatisé afin de pouvoir ainsi l'utiliser pour la sélection des échantillons.

où P_1 est la probabilité de sélection d'un thana donné; a est le nombre de thanas sélectionnés (600 dans cet exemple); et m_t est le nombre de ménages ruraux[12] du $t^{ième}$ thana, conformément au cadre d'échantillonnage utilisé (par exemple le dernier recensement de la population).

101. Le facteur $\sum m_t$ est le nombre total de ménages ruraux de tous les thanas du pays. Il y a lieu de noter que le nombre de thanas effectivement sélectionnés peut être inférieur à 600. Tel peut être le cas lorsqu'un ou plusieurs thanas sont sélectionnés deux fois, ce qui est possible dès lors que la taille d'un thana dépasse l'intervalle d'échantillonnage. L'intervalle d'échantillonnage pour la sélection des thanas est donné par la formule $\sum m_t \div a$.

Ainsi, si l'intervalle d'échantillonnage est, par exemple, de 12 500 et si le thana contient 13 800 ménages, il sera automatiquement sélectionné une fois et aura une chance égale à 1 300/12 500 d'être sélectionné deux fois (le numérateur est égal à 13 800 – 12 500).

3.6.2.2. Deuxième phase de la sélection : districts

102. Lors de la deuxième phase, il est sélectionné un district dans chaque thana pris comme échantillon, dans ce cas également sur la base *d'une probabilité proportionnelle à la taille.* Dans la pratique, il est établi une liste de tous les districts du thana sélectionné, il est fait le total de leur taille, m_u, et il est choisi au hasard un nombre compris entre 1 et m_t, mesure de la taille du thana retenu comme échantillon. Le district dont la taille est la plus réduite par rapport au chiffre sélectionné au hasard permet de déterminer le district à sélectionner (ou bien celui-ci est identifié au moyen d'une convention équivalente). Si un thana a été sélectionné plus d'une fois lors de la première fois, il sera alors sélectionné le même nombre de districts. La probabilité à cette deuxième phase est donnée par l'équation :

$$P_2 = (1)\left(m_u\right)\Big/ m_t \qquad (3.12)$$

où P_2 est la probabilité de sélectionner un district déterminé dans le thana pris comme échantillon; (1) signifie qu'il n'est sélectionné qu'un seul district; et est le nombre de ménages dans le $u^{ième}$ district, conformément au cadre d'échantillonnage.

3.6.2.3. Troisième phase de sélection : villages

103. Lors de la troisième phase, il est sélectionné un village sur la base d'une *probabilité proportionnelle à la taille* dans chaque district retenu. La probabilité à cette troisième phase est donnée par l'équation :

$$P_3 = (1)\left(m_v\right)\Big/ m_u \qquad (3.13)$$

où P_3 est la probabilité de sélectionner un district déterminé dans le thana pris comme échantillon; (1) signifie qu'il n'est sélectionné qu'un seul village; et m_v est le nombre de ménages dans le $v^{ième}$ village, conformément au cadre d'échantillonnage.

12 Ceci est la mesure de la taille et l'on peut retenir plutôt la population du thana, à condition que le chiffre retenu cadre avec toutes les mesures de taille à toutes les phases.

3.6.2.4. *Quatrième phase de sélection : ménages*

104. Lors de la quatrième phase, il sera pris pour hypothèse que l'on dispose pour chaque village sélectionné d'une liste cadre de ménages de sorte que l'échantillon de ménages pourra être sélectionné systématiquement. Il est sélectionné dans chaque village pris comme échantillon un nombre fixe de ménages, qui est la taille prédéterminée de la grappe. La probabilité à cette quatrième phase est donnée par l'équation :

$$P_4 = (b)/m_v \qquad (3.14)$$

où P_4 est la probabilité de sélection d'un ménage donné dans le village pris comme échantillon; et b est le nombre fixe de ménages sélectionnés dans chaque village.

3.6.2.5. *Probabilité globale de la sélection*

105. La probabilité globale, qui est le produit des probabilités à chaque phase, est donnée par la formule :

$$P = P_1\, P_2\, P_3\, P_4 \qquad (3.15)$$

Par substitution, l'on obtient :

$$P = \left[(am_t)/\sum m_t\right]\left[(1)(m_u)/m_t\right]\left[(1)(m_v)/m_u\right]\left[b/m_v\right] = \left[(a)(b)\right]/\sum m_t \qquad (3.16)$$

106. Il y a lieu de noter que P_2 et P_3 s'annulent totalement, ce qui démontre le caractère factice du processus de sélection en « quatre » phases. Ainsi, les thanas et les districts, bien que physiquement « sélectionnés », ont simplement pour but de déterminer où se trouvent les villages pris comme échantillon.

3.6.3. La conception en deux phases

107. L'utilisation d'une méthode de conception de l'échantillon en deux phases dans les pays en développement est une question qui a beaucoup retenu l'attention ces derniers temps. C'est la conception privilégiée pour les Enquêtes en grappes à indicateurs multiples (MICS) effectuées par le Fonds des Nations Unies pour l'enfance (UNICEF) dans plus d'une centaine de pays depuis le milieu des années 90. Ce type de conception est aussi celui qui est le plus généralement utilisé pour les enquêtes démographiques et sanitaires (DHS).

108. Habituellement, une conception en deux phases consiste simplement à sélectionner, sur la base d'une probabilité proportionnelle à la taille, un échantillon de plusieurs centaines de zones géographiques, dûment stratifiées, lors de la première phase. Il peut être établi alors une liste à jour des ménages, selon la disponibilité d'information concernant leur adresse et/ou leur emplacement, et selon que cette information est ou non à jour. Il est alors sélectionné un échantillon systématique d'un nombre fixe de ménages, ce qui constitue la deuxième phase. Les zones géographiques, communément appelées « grappes », sont habituellement des villages ou des zones d'énumération du recensement dans les régions rurales ou des pâtés de maisons dans les régions urbaines.

109. Cette conception en deux phases est attrayante à bien des égards, mais surtout en raison de sa simplicité. *Il est toujours avantageux, lors de la conception de l'échantillon, de rechercher la simplicité plutôt que la complexité afin de réduire le risque d'erreurs autres que des erreurs d'échantillonnage lors de l'application de l'échantillon*. La conception en deux phases présente des caractéristiques utiles qui en font une méthode relativement simple et attrayante. Par exemple :

- Comme on l'a vu, la conception de l'échantillon est autopondérée (tous les ménages de l'échantillon sont sélectionnés avec la même probabilité), ou à peu près autopondérée (voir les sections 3.7.1 et 3.7.2 pour la distinction à établir entre les échantillons sélectionnés sur la base d'une probabilité proportionnelle à la taille et les ménages sélectionnés sur la base d'une probabilité proportionnelle à la taille estimative).
- Les grappes définies en termes de ZE ou de pâtés de maisons sont d'une taille commode (pas trop grande) dans la plupart des pays, surtout s'il doit être établi une nouvelle liste des ménages avant la dernière phase de la sélection.
- Il est habituellement établi une carte des ZE, des pâtés de maisons et de la plupart des villages, soit en vue des recensements, soit à d'autres fins, avec des limites bien définies.

3.7. Échantillonnage sur la base d'une probabilité proportionnelle à la taille et d'une probabilité proportionnelle à la taille estimative

110. L'on a vu dans la section 3.5 un exemple d'échantillonnage sur la base d'une probabilité proportionnelle à la taille, méthode importante pour la sélection des grappes devant constituer l'échantillon. La présente section examine plus en détail l'échantillonnage fondé sur la *probabilité proportionnelle* à la taille.

3.7.1. Échantillonnage sur la base d'une probabilité proportionnelle à la taille

111. En utilisant des méthodes d'échantillonnage fondé sur la *probabilité proportionnelle* à la taille, l'enquêteur peut mieux contrôler, dans les enquêtes en grappes, la taille qu'aura en définitive l'échantillon. Lorsque les grappes sont de même taille ou à peu près, il n'y a guère intérêt à utiliser des méthodes d'échantillonnage sur la base d'une *probabilité proportionnelle*. Supposons par exemple que tous les pâtés de maisons d'une ville contiennent exactement 100 ménages et que l'on veuille constituer un échantillon de 1 000 ménages répartis sur un échantillon de 50 pâtés de maisons. Le plan d'échantillonnage évident consisterait à sélectionner un échantillon aléatoire simple de 50 pâtés de maisons, c'est-à-dire un échantillon fondé sur une probabilité égale, puis à sélectionner systématiquement 1 ménage sur 5 dans chaque pâté de maisons (ce qui est également un échantillon fondé sur une probabilité égale). Il en résulterait un échantillon comportant précisément 20 ménages par pâté de maisons, soit 1 000 ménages en tout. L'équation de sélection serait alors :

$$p = (50/M)(1/5)$$

où p est la probabilité de sélection d'un ménage; $(50/M)$ est la probabilité de sélection d'un pâté de maisons; M est le nombre total de pâtés de maisons que comporte la ville; et $(1/5)$ est la probabilité de sélection d'un ménage dans un pâté de maisons donné.

112. P se ramène à $10/M$. Comme M est une constante, la probabilité globale de sélection de chaque ménage est égale à 10 divisé par le nombre de pâtés de maisons, M.

113. Dans la réalité, cependant, les pâtés de maisons ou autres unités géographiques pouvant être utilisés comme grappes pour une enquête sur les ménages ont rarement une taille uniforme. Dans l'exemple ci-après, la taille pourrait varier entre, par exemple, 25 et 200. Un échantillon de pâtés de maisons sélectionné sur la base d'une probabilité égale peut entraîner par une « malchance » une sélection de pâtés de maisons pour la plupart de petite taille ou pour la plupart de grande taille. En pareil cas, l'on aurait une taille globale très différente des 1 000 ménages recherchés. Pour réduire le risque que les tailles de l'échantillon varient beaucoup, l'on peut notamment créer des strates fondées sur la taille des grappes et sélectionner un échantillon dans chaque strate. Cette méthode n'est pas généralement recommandée car elle peut réduire ou compliquer l'utilisation, lors de la conception de l'échantillon, d'autres facteurs de stratification. Un échantillonnage sur la base d'une probabilité proportionnelle à la taille est la solution privilégiée car elle permet de mieux contrôler la taille de l'échantillon ultime sans devoir procéder à une stratification par taille.

114. Pour illustrer un échantillonnage sur la base d'une probabilité proportionnelle à la taille, l'on prend pour point de départ la formule de sélection mentionnée ci-dessus mais exprimée de manière plus formelle pour une conception en deux phases[13] comme suit :

$$P(\alpha\beta) = P(\alpha)P(\beta|\alpha), \qquad (3.17)$$

où $P(\alpha\beta)$ est la probabilité de sélection du ménage β dans la grappe α; $P(\alpha)$ est la probabilité de sélection de la grappe α, $P(\beta|\alpha)$ est la probabilité conditionnelle de sélection du ménage β à la deuxième phase, étant donné que la grappe α était sélectionnée lors de la première phase.

115. Pour déterminer la taille globale de l'échantillon en termes de nombre de ménages, il nous faut un échantillon de n ménages sélectionnés sur la base d'une *probabilité égale,* sur une population de N ménages. Ainsi, le taux global d'échantillonnage est n/N, ce qui est égal à $P(\alpha\beta)$, comme défini ci-après. En outre, si le nombre de grappes à interroger est spécifié comme étant a, idéalement, il nous faudra sélectionner b ménages dans chaque grappe, quelle que soit la taille des grappes sélectionnées. Si nous définissons m_i comme étant la taille de la $i^{ième}$ grappe, il faudra que $P(\beta|\alpha)$ soit égal à b/m_i. Donc,

$$P(\alpha\beta) = [P(\alpha)]\left[b/m_i\right].$$

Comme $n = ab$, nous avons

$$ab/N = [P(\alpha)]\left[b/m_i\right].$$

En résolvant la dernière équation pour $P(\alpha)$, nous obtenons

$$P(\alpha) = (a)(m_i)/N \qquad (3.18)$$

116. Il y a lieu de noter que $N = \sum m_i$, de sorte que la probabilité de sélection d'une grappe est proportionnelle à sa taille. L'équation de sélection pour un échantillon d'unités primaires établi sur la base d'une *probabilité proportionnelle à la taille* dans lequel les unités ultimes sont néanmoins sélectionnées sur la base d'une probabilité égale est par conséquent :

13 Voir Kalton (1983, p. 38 à 47) pour le développement de cette notation et une discussion plus détaillée de l'échantillonnage sur la base d'une probabilité proportionnelle à la taille.

$$P(\alpha\beta) = \left[(a)(m_i)/\sum m_i\right]\left[b/m_i\right] \quad (3.19)$$

$$= \left[(ab)/\sum m_i\right] \quad (3.20)$$

117. La conception de l'échantillon ainsi obtenu est autopondérée, comme le montre l'équation (3.19), étant donné que tous les termes de l'équation sont constants; il ne faut pas perdre de vue que si m_i est une variable, la somme $\sum m_i$ est une constante égale à N. La figure 3.2 ci-après donne un exemple de sélection d'un échantillon de grappes sur la base d'une *probabilité proportionnelle à la taille*.

118. En ce qui concerne la sélection physique de l'échantillon, il y a lieu de noter que dans la figure 3.2, l'intervalle d'échantillonnage, *I,* est ajouté successivement à l'abord aléatoire (RS) sept fois (ou a – 1 fois, où a est le nombre de grappes à sélectionner). Les nombres de sélection sont alors de 311,2 (qui est le *RS*), 878,8, 1 446,4, 2 014, 2 581,6, 3 149,2, 3 716,8 et 4 284,4. La grappe sélectionnée pour ces huit nombres de sélection est, dans chaque cas, celle dont la taille cumulée est égale ou supérieure au nombre correspondant. Ainsi, l'on sélectionne la grappe 03 parce que 377 est le nombre global le plus petit tout en étant égal ou supérieur à 311,2, et la grappe 26 est sélectionnée parce que 3 744 est le nombre global le plus petit tout en étant égal ou supérieur à 3 716,8.

119. Bien que cela ne soit pas démontré de façon concluante par l'exemple (car il n'est sélectionné que huit grappes), l'échantillonnage sur la base d'une probabilité proportionnelle à la taille tend à déboucher sur des grappes de dimensions plus grandes plutôt que plus petites. Cela est peut-être évident étant donné que la formule (3.17) montre que la probabilité de sélection d'une grappe est proportionnelle à sa taille; ainsi une grappe contenant 200 ménages a deux fois plus de chances d'être sélectionnée qu'une grappe de 100 ménages. Il y a donc lieu de noter que la même grappe peut être sélectionnée plus d'une fois si sa taille dépasse l'intervalle d'échantillonnage, *I*. Cependant, aucune des grappes de la figure ne répond à cette condition; si tel est le cas, cependant, le nombre de ménages à sélectionner dans une telle grappe serait deux fois plus élevé si la grappe est sélectionnée deux fois, trois fois plus élevé si elle l'est trois fois, et ainsi de suite.

3.7.2. Échantillonnage sur la base d'une probabilité proportionnelle à la taille estimative

120. La méthode d'échantillonnage fondée sur la *probabilité proportionnelle à la taille* décrite dans la section précédente est pour une large part idéale et peut, le plus souvent, être difficile à appliquer dans la pratique. En effet, la taille utilisée pour établir la probabilité de sélection de la grappe au niveau primaire n'est souvent pas la taille *effective* de la grappe lorsque l'échantillon de ménages est sélectionné lors de la phase secondaire.

121. Dans les enquêtes sur les ménages, la taille généralement adoptée pour la sélection primaire des unités primaires d'échantillonnage ou grappes est le nombre de ménages (ou les effectifs de la population) provenant du dernier recensement. Même si celui-ci est très récent, le nombre effectif de ménages au moment de l'enquête sera sans doute différent, même si l'écart est modeste. Il y a cependant une exception lorsque la sélection des ménages lors de la phase secondaire est faite directement à partir du même cadre que celui qui est utilisé pour déterminer la taille (pour une discussion plus détaillée des cadres d'échantillonnage, voir le chapitre 4).

Figure 3.2
Exemple de sélection systématique de grappes sur la base d'une probabilité proportionnelle à la taille

Nombre de grappes/UPE	Taille *(nombre de ménages)*	Chiffre cumulé	Sélection de l'échantillon
001	215	215	
002	73	288	
003	89	377	311,2
004	231	608	
005	120	728	
006	58	786	
007	99	885	878,8
008	165	1 050	
009	195	1 245	
010	202	1 447	1 446,4
011	77	1 524	
012	59	1 583	
013	245	1 828	
014	171	1 999	
015	99	2 098	2 014,0
016	88	2 186	
017	124	2 310	
018	78	2 388	
019	89	2 477	
020	60	2 537	
021	222	2 759	2 581,6
022	137	2 896	
023	199	3 095	
024	210	3 305	3 149,2
025	165	3 470	
026	274	3 744	3 716,8
027	209	3 953	
028	230	4 183	
029	67	4 250	
030	72	4 322	4 284,4
031	108	4 430	
032	111	4 541	

Instructions : Sélectionner 8 UPE (grappes) sur les 32 de l'univers sur la base d'une probabilité proportionnelle à la taille; l'intervalle de sélection (*I*) est par conséquent égal à 4 541/8, OU 567,6, 4 541 étant la taille globale cumulée pour toutes les grappes et 8 le nombre de grappes à sélectionner; l'abord aléatoire (RS) est le chiffre compris entre 0,1 et 567,6 choisi au hasard sur un tableau de chiffres aléatoires; dans cet exemple, RS = 311,2.

Exemple

Supposons qu'une enquête sur les ménages soit réalisée trois mois après la fin du recensement de la population. Plutôt que d'établir une nouvelle liste des ménages des grappes sélectionnées, il est décidé de se référer à la liste des ménages établie lors du recensement pour la phase secondaire, l'idée étant que l'on peut tenir pour acquis que la liste établie lors du recensement est, à toutes fins utiles, à jour et exacte. Au niveau primaire, il est sélectionné un échantillon de villages, la taille de chaque village étant déterminée sur la base du dénombrement réalisé lors du recensement. Pour chaque village faisant partie de l'échantillon, la taille m_i est identique au nombre effectif de ménages parmi lesquels l'échantillon doit être sélectionné. Ainsi, si l'on sélectionne un village A qui comporte 235 ménages, selon le recensement, la liste à partir de laquelle sera sélectionné l'échantillon de ménages aux fins de l'enquête comportera également 235 ménages.

122. Il arrive cependant souvent que l'enquête soit réalisée plusieurs mois, voire des années, après le recensement (voir le chapitre 4 pour une discussion plus détaillée de la mise à jour des cadres d'échantillonnage). Cela étant, il est souvent décidé d'organiser sur le terrain l'établissement d'une nouvelle liste des ménages faisant partie des grappes sélectionnées pour inclusion dans l'échantillon au niveau primaire. Un échantillon de ménages est alors sélectionné aux fins de l'enquête sur la base de cette nouvelle liste.

123. La taille, m_i, utilisée pour sélectionner la grappe, est le nombre de ménages identifiés lors du recensement, comme indiqué dans l'exemple ci-dessus. Toutefois, la liste effective à partir de laquelle sera sélectionné l'échantillon de ménages sera différente. Elle aura notamment, dans une certaine mesure, une taille différente, selon l'intervalle qui s'est écoulé depuis le recensement et l'établissement de la nouvelle liste. Il y aura des différences en raison des migrations, de la construction de nouveaux logements, de la destruction de maisons existantes, de la constitution de nouveaux ménages à la suite de mariages (parfois dans la même habitation que les parents) et de décès. Lorsque l'échantillon est sélectionné sur la base d'une *probabilité proportionnelle à la taille estimative,* l'équation qui donne la probabilité de sélection est :

$$P(\alpha\beta) = \left[(a)(m_i)/\sum m_i\right]\left[b/m_i'\right] \qquad \textbf{(3.21)}$$

où m_i' est le nombre de ménages figurant sur la liste, et les autres termes sont définis comme précédemment.

124. Comme m_i' et m_i seront sans doute des valeurs différentes pour la plupart des grappes, sinon toutes, le calcul de la probabilité de sélection (et par conséquent de la pondération, c'est-à-dire l'inverse de la probabilité) doit tenir compte de la différence. Comme le montre l'équation 3.20, chaque grappe aura une pondération différente, ce qui rend impossible une conception autopondérée.

125. L'on pourra éviter d'introduire des distorsions dans les estimations résultant de l'enquête en utilisant les pondérations exactes de nature à compenser la différence entre les tailles identifiées lors du recensement et de l'enquête. Si les pondérations ne sont pas ajustées en conséquence, l'on risque d'avoir des distorsions qui seront indubitablement plus sérieuses à mesure que s'allonge l'intervalle entre le recensement et l'enquête. Il y a lieu de noter toutefois que, lorsque les différences entre m_i' et m_i sont mineures, l'échantillon est virtuellement autopondéré et qu'il peut être prudent, dans certaines

circonstances[14], de générer des estimations sans pondération, car les distorsions seraient négligeables. Avant de suivre cette démarche, cependant, il est essentiel d'examiner m_i' et m_i, grappe par grappe, pour s'assurer que les différences sont effectivement mineures.

126. L'on peut suivre une autre stratégie pour sélectionner les ménages lors de la dernière phase lorsque l'on a recours à un échantillonnage avec une probabilité proportionnelle à la taille estimative, lorsque l'échantillon est effectivement autopondéré. Cette stratégie consiste à sélectionner les ménages à un taux variable à l'intérieur de chaque grappe, selon sa taille effective (comme on le verra dans la prochaine section).

3.8. Options pouvant être envisagées pour l'échantillonnage

127. La présente section évoque certaines des options, et elles sont nombreuses, qui peuvent être envisagées pour concevoir un échantillon approprié en vue d'une enquête de caractère général, l'accent étant mis surtout sur les stratégies à suivre à l'avant-dernière et à la dernière phases de la sélection, car ce sont celles auxquelles se présentent plusieurs possibilités. Elle examine le choix entre la sélection de grappes sur la base d'une *probabilité égale* ou *d'une probabilité proportionnelle à la taille* à l'avant-dernière phase ainsi qu'entre un échantillonnage sur la base d'un taux fixe par opposition à une taille fixe lors de la dernière phase et, dans une certaine mesure, elle résume les sections précédentes en ce qui concerne les questions liées au contrôle de la taille de l'échantillon, à la conception d'un échantillon autopondéré ou non autopondéré ainsi que d'autres questions comme la charge de travail de l'enquêteur. En outre, elle passe en revue des types spécifiques de conceptions qui sont largement utilisées aujourd'hui, comme celles qui sont à la base des enquêtes démographiques et sanitaires et de l'Enquête en grappes à indicateurs multiples réalisée par l'UNICEF. Ces conceptions sont d'autres options qui méritent d'être prises en compte, et tel est notamment le cas de l'utilisation de grappes compactes (globales) et non compactes.

3.8.1. Échantillonnage à probabilité égale, échantillonnage à probabilité proportionnelle à la taille et échantillonnage sur la base d'un taux fixe et d'une taille fixe

128. Le tableau 3.4 illustre le cadre qui peut être retenu pour discuter des procédures, des conditions, des avantages et des limitations des divers plans d'échantillonnage.

Tableau 3.4
Divers plans d'échantillonnage : deux dernières phases de la sélection

Sélection des avant-dernières unités	Grappes de taille fixe (nombre de ménages)	Taux fixe de sélection dans chaque grappe
Probabilité proportionnelle à la taille	Plan 1	Plan 2 [pas recommandé]
Probabilité proportionnelle à la taille estimative	Plan 3	Plan 4 [pas recommandé]
Probabilité égale	Plan 5	Plan 6

14 Dans le cas d'enquêtes visant à établir seulement des proportions, des taux ou des ratios, ce serait là une stratégie appropriée; cependant, lorsque l'on cherche à obtenir des totaux ou des chiffres absolus, il faut utiliser une pondération, sans égard à la question de savoir si l'échantillon est autopondéré, presque autopondéré ou non autopondéré.

129. Nous avons vu comment un échantillonnage, sur la base d'une probabilité proportionnelle à la taille, des unités primaires d'échantillonnage ou grappes constitue un moyen de contrôler plus exactement la taille qu'aura en définitive l'échantillon que s'il est utilisé une méthode d'échantillonnage fondée sur une *probabilité égale,* et c'est là que réside son principale avantage, surtout si les grappes sont très variables pour ce qui est du nombre de ménages que chacune d'elles contient. Il est important de contrôler la taille de l'échantillon non seulement en raison des incidences qu'elle a sur le coût de l'opération mais aussi pour que les responsables de l'enquête puissent planifier comme il convient les entrevues auxquelles devront procéder les enquêteurs. La sélection d'un échantillon sur la base d'une *probabilité égale,* en revanche, est plus simple que si elle est effectuée sur la base d'une *probabilité proportionnelle à la taille,* et il est logique d'utiliser cette méthode lorsque la taille des grappes est à peu près égale ou varie peu. Dans la pratique, il faut utiliser un échantillonnage sur la base d'une probabilité proportionnelle à la taille estimative plutôt que proportionnelle à la taille dans tous les cas où la taille effective est autre que celle donnée par le cadre d'échantillonnage.

130. La sélection d'un nombre fixe de ménages dans chaque grappe a deux avantages très importants : premièrement, la taille de l'échantillon est contrôlée avec précision et, deuxièmement, cette méthode permet au responsable de l'enquête de délimiter avec précision la charge de travail des enquêteurs et de l'égaliser s'il y a lieu. L'échantillonnage sur la base d'une taille fixe, cependant, est une méthode assez complexe dans la mesure où il faut calculer pour chaque grappe les différences d'intervalle d'échantillonnage. Or, appliquer des intervalles d'échantillonnage différents peut susciter des confusions et être à l'origine d'erreurs. Il y a cependant un contrôle incorporé de la qualité étant donné que le nombre de ménages à sélectionner est connu à l'avance. Il n'en demeure pas moins que, du fait de sa complexité, cette méthode peut être inefficace en raison du temps qu'il faut consacrer à la correction des erreurs de sélection.

131. L'échantillonnage sur la base d'une taille fixe exige, par définition, une liste des ménages permettant de désigner et d'identifier ceux qui seront sélectionnés. Le plus souvent, il s'agit d'une liste établie lors de la préparation de l'enquête, et il est bon de veiller à ce que la sélection des ménages se fasse sur une base centralisée et qu'elle soit effectuée par une personne autre que celle qui a établi la liste afin de réduire au minimum le risque de distorsion dans la procédure de sélection.

132. Il peut également être sélectionné un échantillon de ménages sur la base d'un taux fixe à l'intérieur de chaque grappe. En pareil cas, la sélection est plus simple et présente moins de risque d'erreur. Un avantage sur le terrain est que l'échantillonnage peut être fait lorsque l'enquêteur mène son quadrillage pour obtenir une liste à jour des ménages. À cette fin, il doit concevoir la liste de manière à ce que celle-ci comporte des lignes prédéterminées pour identifier les ménages qui feront partie de l'échantillon. Le fait que la liste peut être établie en même temps qu'il est procédé à l'échantillonnage a des avantages évidents en termes de coût, mais cette méthode comporte également certaines limitations importantes.

133. Une des limitations que comporte l'échantillonnage sur la base d'un taux fixe est qu'il est difficile de contrôler la taille de l'échantillon ou la charge de travail de l'enquêteur, à moins que la *taille* de chaque grappe soit approximativement la même. Une autre limitation, plus grave, est que lorsque les enquêteurs se voient confier le soin de sélectionner effectivement les ménages qui feront partie de l'échantillon, c'est-à-dire d'identifier ceux qui figureront sur la liste, des distorsions se trouvent fréquemment introduites dans la sélection. D'innombrables études ont montré que les ménages sélectionnés, lorsque le choix est effectué par les enquêteurs, tendent à être moins nombreux, ce qui

porte à penser que les enquêteurs, consciemment ou non, choisissent des ménages comportant moins de déclarants afin de réduire leur charge de travail.

134. Un échantillon est autopondéré lorsqu'il ne dépend pas des différentes méthodes d'échantillonnage suivies lors de chaque phase. Ainsi, une conception en deux phases prévoyant un échantillonnage de grappes sur la base d'une *probabilité proportionnelle à la taille* et un échantillonnage des ménages sur la base d'une taille fixe est autopondéré tandis que la combinaison d'une *probabilité proportionnelle à la taille* et d'un taux fixe ne l'est pas. La discussion qui suit indique quels sont ceux des plans indiqués au tableau 3.4 qui sont autopondérés.

3.8.1.1. Plan 1 : probabilité proportionnelle à la taille, grappes de taille fixe

Conditions

- Taille variable pour l'univers de grappes.
- Les ménages sont sélectionnés sur les mêmes listes (par exemple liste des ménages établie lors du recensement) que celles qui sont utilisées pour la détermination de la *taille.*

Avantages

- Maîtrise de la taille globale de l'échantillon et par conséquent des coûts.
- Contrôle de la charge de travail des enquêteurs.
- Autopondération

Limitations

- *Un échantillonnage sur la base d'une probabilité proportionnelle à la taille est une méthode beaucoup plus difficile à appliquer que celle d'un échantillonnage fondé sur une probabilité égale.*
- Taux de sélection différents pour le choix des ménages dans chaque grappe, ce qui crée un risque d'erreur.

3.8.1.2. Plan 2 : probabilité proportionnelle à la taille, taux fixe

135. Il est difficile de prévoir une situation dans laquelle ce type de conception serait utilisé. Si les grappes sont de tailles variables, il y aura lieu d'utiliser un échantillonnage sur la base d'une probabilité proportionnelle à la taille en même temps que des grappes de taille fixe. Si les grappes sont de tailles à peu près égales, un échantillonnage sur la base d'un taux fixe sera approprié, mais les grappes elles-mêmes devront être sélectionnées sur la base d'un échantillonnage à probabilité égale.

3.8.1.3. Plan 3 : probabilité proportionnelle à la taille estimative, taille fixe des grappes

Conditions

- *Taille* variable pour l'univers de grappes.
- Les ménages sont sélectionnés sur les mêmes listes (par exemple liste des ménages établie lors du recensement) à partir de nouvelles listes mettant à jour celles du cadre d'échantillonnage afin de déterminer la *taille* initiale.

Avantages

- Maîtrise de la taille globale de l'échantillon et par conséquent des coûts.
- Contrôle de la charge de travail des enquêteurs.
- Méthode plus exacte que celle de la probabilité proportionnelle à la taille pour un cadre donné car les listes de ménages sont à jour.

Limitations

- Un échantillonnage sur la base d'une probabilité proportionnelle à la taille estimative est une méthode beaucoup plus difficile à appliquer que celle d'un échantillonnage fondé sur une probabilité égale.
- Taux de sélection différents pour le choix des ménages dans chaque grappe, ce qui crée un risque d'erreur.
- Pas autopondérée.

3.8.1.4. Plan 4 : probabilité égale, taux fixe

136. L'on ne voit pas de situation dans laquelle le plan 4 pourrait être utilisé, pour les raisons indiquées ci-dessus pour le plan 2.

3.8.1.5. Plan 5 : probabilité égale, grappe de taille fixe

Conditions

- Les *tailles* sont approximativement égales ou varient très peu pour l'univers de grappes.

Avantages

- Maîtrise de la taille de l'échantillon global (mais un peu moins que dans le cas du plan 1 et par conséquent des coûts.
- Contrôle de la charge de travail des enquêteurs mais, dans ce cas également, un peu moins que dans le cas du plan 1.
- La méthode de la probabilité égale est plus facile à appliquer que celle de la probabilité proportionnelle à la taille ou de la probabilité proportionnelle à la taille estimative.

Limitations

- Taux de sélection différents pour le choix des ménages dans chaque grappe, ce qui introduit un risque d'erreur.
- Pas autopondérée.

3.8.1.6. Plan 6 : probabilité égale, taux fixe

Conditions

- Les *tailles* sont presque égales pour l'univers de grappes.

Avantages

- Autopondération.
- Il est très simple de sélectionner l'échantillon lors des deux phases.

Limitations

- Contrôle réduit de la taille de l'échantillon global, ce qui a des incidences sur les coûts et la fiabilité, surtout si la taille actuelle est très différente de la taille prévue par le cadre, et moindre fiabilité si l'échantillon est beaucoup plus petit que prévu.
- Peu de contrôle sur la charge de travail des enquêteurs.

3.8.2. Enquête démographique et sanitaire (DHS)

137. Bien que l'enquête démographique et sanitaire (DHS) soit axée sur les femmes en âge de procréer, la conception de l'échantillon est appropriée pour des enquêtes de caractère général.

138. Les enquêtes démographiques et sanitaires, qui sont courantes dans des dizaines de pays en développement depuis 1984, encouragent l'utilisation de la *conception du segment type*[15], en raison de sa commodité et de sa facilité d'application. Un segment type est défini en termes de taille et comprend habituellement 500 personnes. Chaque circonscription géographique du pays qui fait partie du cadre d'échantillonnage se voit assigner une taille calculée comme étant sa population divisée par 500 (ou par la taille du segment type fixée pour le pays en question). Le résultat, arrondi au chiffre entier le plus proche, est le nombre de segments types que comporte l'unité d'échantillonnage.

139. Il est sélectionné sur la base d'une probabilité proportionnelle à la taille un échantillon de zones dont la taille est le nombre de segments types. Comme les unités utilisées à cette phase de l'échantillonnage sont habituellement des zones d'énumération, des pâtés de maisons ou des villages, la taille est, pour de fortes proportions d'entre eux, égale à un ou deux. Dans le cas d'unités dont la *taille* dépasse un, il est établi une carte des segments géographiques de sorte que leur nombre soit égal à la taille voulue. Ainsi, une zone de *taille* 3 sera subdivisée en trois segments de taille à peu près égale, dans la mesure où les limites naturelles le permettent, de sorte que chaque segment comporte non pas une taille géographique égale mais un nombre égal de personnes.

140. Chaque unité de *taille* 1 fait automatiquement partie de l'échantillon et, dans chacune des autres, il est sélectionné un segment au hasard sur la base d'un échantillonnage à probabilité égale. Tous les segments inclus dans l'échantillon, y compris ceux qui sont choisis automatiquement, font alors l'objet d'un quadrillage pour obtenir une liste à jour des ménages. Une fraction fixe (taux fixe) des ménages est sélectionnée systématiquement dans chaque « grappe » aux fins des entrevues. Comme les segments sont tous à peu près de même taille, la procédure d'échantillonnage donne un échantillon de segments et de ménages sélectionnés en deux phases sur la base d'une probabilité égale.

141. Le segment type utilisé pour la DHS ressemble au plan 6 ci-dessus : sélection des grappes sur la base d'une probabilité égale et sélection des ménages sur la base d'un taux fixe dans les grappes faisant partie de l'échantillon (également par probabilité égale). Cependant, la procédure faisant appel à des segments types permet d'éviter les sérieuses limitations relevées ci-dessus dans le cadre du

15 La méthode du segment type a été utilisée aussi dans le cadre du Programme panarabe d'enquêtes sur le développement des enfants mené pendant les années 80 et 90; voir Ligue des États arabes (1990).

plan 6 : la taille de l'échantillon global est contrôlée de façon presque tout à fait précise, et tel est le cas aussi de la charge de travail des enquêteurs.

142. L'un des principaux avantages d'une conception faisant appel à des segments types est que le travail que représente l'établissement de listes à l'avant-dernière phase de la sélection se trouve considérablement réduit. Pour chaque unité composée de *s* segments, la charge de travail se trouve ramenée à 1/*s* (lorsqu'il n'y a qu'un seul segment, la charge de travail ne diminue pas). Par exemple, si une unité donnée contient quatre segments, le travail que suppose l'établissement d'une liste n'est que le quart de ce qu'il aurait été s'il avait fallu établir une liste pour l'ensemble de la zone. Cela permet simultanément de réduire les coûts de la préparation de l'échantillon.

143. Si l'établissement de listes coûte moins cher, la réduction ne va pas sans prix. L'une des limitations de la méthode du segment type est qu'il faut établir des cartes pour les segments de taille supérieure à 1, ce qui peut être long et coûteux, exige une formation approfondie et entraîne des risques d'erreur. Le fait que, fréquemment, les limites naturelles sont mal définies empêche de délimiter de façon raisonnable les segments faisant partie de l'unité géographique considérée. Aussi est-il difficile pour les enquêteurs qui doivent ensuite se rendre dans ce segment de déterminer avec précision où se trouvent les ménages sélectionnés. Ce problème peut cependant être quelque peu atténué en indiquant le nom du chef de ménage lors de l'établissement de la liste, auquel cas des limites mal tracées suscitent moins de difficultés.

3.8.3. Échantillon en grappes modifié : enquêtes en grappes à indicateurs multiples (MICS)

144. Les praticiens se plaignent souvent de l'investissement de temps et d'argent qu'exige l'établissement de listes des ménages des grappes sélectionnés lors de l'avant-dernière phase. Il faut, dans la plupart des enquêtes, établir des listes — y compris, comme on l'a dit, dans le cas de la méthode du segment type de la DHS — pour obtenir une liste à jour des ménages parmi lesquels seront sélectionnés ceux qu'il faudra interroger. Cela est particulièrement important lorsque le cadre d'échantillonnage remonte à plus d'un an. *L'établissement de listes suppose un travail long et coûteux qui est fréquemment négligé aussi bien lors de l'élaboration du budget que lors de la programmation de l'enquête.* Il faut, pour établir les listes, prévoir des visites sur le terrain indépendamment de celles qu'exigeront les entrevues. De plus, il arrive fréquemment que le ratio de ménages à indiquer sur la liste soit 5 ou 10 fois supérieur au nombre de ménages qui seront sélectionnés. Supposons par exemple qu'il soit prévu de sélectionner 300 UPE, en grappes de 25 ménages, soit en tout 7 500 ménages à interroger. Si l'UPE moyenne à l'avant-dernière phase contient 150 ménages, il faudra établir une liste de 45 000 ménages.

145. La stratégie d'échantillonnage utilisée pour l'enquête en grappes réalisée au sujet du Programme élargi de vaccination (PEV) par l'Organisation mondiale de la Santé (OMS) en 1991 a été élaborée par les Centers for Disease Control et l'OMS, en partie pour éviter l'investissement de temps et d'argent qu'exige l'établissement de listes. L'enquête en grappes sur le PEV, qui a pour but d'évaluer la couverture des programmes de vaccination des enfants, a été largement utilisée dans des dizaines de pays en développement depuis plus de 20 ans. Une importante question statistique (Turner, Magnani et Shuaib, 1996) a trait à la méthode d'échantillonnage. La méthode d'enquête en grappes utilise un échantillon par quotas lors de la deuxième phase de la sélection, alors même que les unités primaires (villages ou quartiers) sont habituellement sélectionnées conformément aux principes de l'échantillonnage probabiliste. La méthode d'échantillonnage par quotas qui est fréquemment uti-

lisée, bien qu'elle comporte des variations, consiste à commencer l'enquête par des entrevues menées en un point central du village sélectionné avant de poursuivre dans une direction déterminée au hasard jusqu'à ce qu'un nombre suffisant de ménages pour constituer le quota aient été interrogés. Selon une variante de ce type d'enquête, les ménages continuent d'être interrogés jusqu'à ce qu'il soit trouvé sept enfants du groupe d'âge cible. S'il n'y a pas de biais délibéré dans l'utilisation de ces types de méthode, nombre de statisticiens l'ont critiquée depuis longtemps, y compris Kalton (1987), Scott (1993) et Bennett (1993). La principale critique est que cette méthode ne permet pas d'obtenir un échantillon probabiliste (voir la section 3.2 sur l'échantillonnage probabiliste par opposition aux autres méthodes d'échantillonnage pour une discussion des raisons pour lesquelles un échantillonnage probabiliste est la méthode recommandée pour les enquêtes sur les ménages).

146. Variante de la méthode d'enquête en grappes sur le PEV, l'enquête en grappes dite modifiée (MCS) a été élaborée pour fournir une stratégie permettant d'éviter l'établissement de listes tout en étant fondée sur un échantillonnage probabiliste. Différentes versions de la MCS ainsi que d'autres types voisins ont été appliqués dans différentes régions du monde dans le cadre des enquêtes en grappes à indicateurs multiples (MICS) menées sous l'égide de l'UNICEF pour suivre la réalisation de certains objectifs concernant la situation des femmes et des enfants arrêtés lors du Sommet mondial pour les enfants (UNICEF, 2000).

147. La conception de la MCS est une stratégie d'échantillonnage minimaliste. Elle utilise une conception simple en deux phases avec une stratification soigneuse et un quadrillage rapide et une segmentation des zones. Il n'est pas établi de liste. Les caractéristiques essentielles de la conception d'un échantillon MCS sont les suivantes :

- Sélection pendant une première phase d'un échantillon d'unités géographiques comme villages ou pâtés de maisons sur la base d'une *probabilité proportionnelle à la taille* ou d'une probabilité égale, selon la variabilité des UPE en termes de taille. Des tailles anciennes peuvent être utilisées même si le cadre de recensement remonte à quelques années, encore que celui-ci doive couvrir intégralement la population visée, qu'elle soit nationale ou localisée.
- Visites de chaque unité d'échantillonnage en vue d'un quadrillage rapide et d'une segmentation des zones sur la base des cartes ou croquis existants, le nombre de segments étant prédéterminé et égal à la taille du recensement divisée par la taille souhaitée (prévue) des grappes. Les segments créés renferment une population dont le nombre est approximativement égal.
- Sélection d'un segment sur la base d'une probabilité égale dans chaque UPE faisant partie de l'échantillon.
- Entrevues avec tous les ménages de chaque segment sélectionné.

148. L'utilisation d'une segmentation sans qu'il soit nécessaire d'établir des listes est le principal avantage de la conception fondée sur la taille, qui diffère de la conception fondée sur le segment type de la DHS (Demographic Health Survey), laquelle exige l'établissement d'une liste pour chaque segment. En outre, l'opération de segmentation compense en partie le fait qu'il est utilisé un cadre qui peut être dépassé. Si cela a l'avantage de produire une estimation dépourvue de distorsion, cette méthode a cependant l'inconvénient qu'il est plus difficile de contrôler la taille de l'échantillon ul-

time étant donné que, s'étant développé, l'un des segments sélectionnés peut être d'une taille bien plus grande que celle qui ressort du cadre.

149. La méthode fondée sur la taille exige cependant l'établissement de cartes, tout comme la méthode du segment type de la DHS, avec toutes les limitations que cela suppose, comme indiqué dans le contexte de la DHS. En outre, la création de petits segments dont la taille correspond à celle des grappes et dont les limites sont définies avec précision peut être difficile lorsqu'il n'existe pas de limites naturelles dans des zones peu étendues. Il y a une dernière limitation : dans la mesure où le segment interrogé est une grappe compacte, tous les ménages sont géographiquement contigus, l'effet de conception est plus marqué en raison d'une corrélation intraclasse relativement élevée que celui que produisent les grappes non compactes utilisées dans la méthode du segment type.

3.9. Options spéciales : échantillons en deux phases et échantillonnage pour l'estimation de tendances

150. La présente section traite de deux aspects particuliers de la conception de l'échantillon dans les enquêtes sur les ménages : *a*) l'échantillonnage en deux phases, la première phase consistant en une brève entrevue visant à identifier, parmi les ménages résidents, les personnes qui font partie de la population cible et la seconde consistant à sélectionner un échantillon de personnes répondant aux critères fixés; et *b*) la méthode d'échantillonnage utilisée lorsqu'une enquête est répétée pour estimer un changement ou une tendance.

3.9.1. Échantillonnage en deux phases

151. Un échantillon d'une conception spéciale est nécessaire pour les enquêtes sur les ménages lorsque l'on ne dispose pas d'informations suffisantes pour sélectionner efficacement un échantillon de la population cible. Tel est généralement le cas lorsque la population cible est une sous-population, souvent rare, dont les membres ne sont présents que dans une faible proportion de ménages. L'on peut en citer comme exemples les membres d'un groupe ethnique particulier, les orphelins et les personnes dont les revenus sont inférieurs ou supérieurs à un niveau déterminé. Une stratification judicieuse peut fréquemment être utilisée pour identifier, par exemple, les zones où se trouvent concentrés le groupe ethnique visé ou les personnes à revenu élevé mais, lorsque ces groupes sont dispersés assez au hasard parmi l'ensemble de la population ou lorsque le groupe cible, comme les orphelins, est rare, une stratégie ne suffit pas et il faut avoir recours à d'autres méthodes pour en établir un échantillon.

152. Une méthode fréquemment utilisée est celle de l'échantillonnage en deux phases, également appelée échantillonnage post-stratifié ou double échantillonnage. Cette méthode comporte quatre étapes :

a) Sélection d'un échantillon « nombreux » de ménages;

b) Brève entrevue de quadrillage pour identifier les ménages parmi lesquels résident des membres de la population cible;

c) Post-stratification de l'échantillon nombreux en deux catégories fondées sur l'entrevue de quadrillage;

d) Sélection d'un sous-échantillon de ménages dans chacune des deux strates en vue d'une deuxième entrevue, plus longue, avec le groupe cible.

153. L'objectif de cette approche en deux phases est de réduire les coûts de l'opération en procédant à un bref entretien de quadrillage avec les membres du premier échantillon, nombreux. Ce bref entretien est suivi d'une entrevue plus détaillée, à une date ultérieure, à laquelle participent seulement les ménages sélectionnés. Pour cette raison, l'échantillon initial est souvent un échantillon choisi à d'autres fins, l'entrevue de quadrillage étant « greffée » sur l'enquête principale. Cette procédure permet ainsi d'allouer la majeure partie des ressources à la deuxième phase et n'exige qu'un modeste budget pour les entretiens requis lors de la phase de quadrillage.

Exemple

Supposons qu'il soit prévu de réaliser une enquête parmi 800 orphelins résidents parmi des ménages de parents survivants ou d'autres membres de la famille (par opposition aux orphelins qui vivent en établissement). Supposons en outre que l'on estime devoir interroger 16 000 ménages pour trouver 800 orphelins, soit environ 1 orphelin pour 20 ménages. En raison du coût que supposerait la conception et l'administration d'une enquête auprès de 16 000 ménages pour procéder à 800 entrevues détaillées seulement, il est décidé de profiter d'une enquête sur la santé de caractère général qui est également prévue. Cette enquête sur la santé doit interroger un échantillon de 20 000 ménages. Les responsables des deux enquêtes conviennent de greffer sur l'enquête sanitaire une seule question, à savoir : y a-t-il dans ce village une personne de moins de 17 ans dont le père, la mère ou les deux parents sont décédés ? Cette question accessoire devrait permettre d'identifier des ménages comportant environ 1 000 orphelins. Le responsable de l'enquête sur les orphelins préparerait alors des entrevues détaillées avec 80 % de ces ménages.

154. L'exemple ci-dessus illustre également *quand* un échantillonnage en deux phases est approprié. Il y a lieu de noter que la taille de l'échantillon visé dans l'exemple est de 800 orphelins seulement, bien que le nombre de ménages nécessaire pour trouver ce nombre d'orphelins soit de 16 000. Ainsi, pour calculer ce dernier chiffre (voir la formule 3.7), le technicien et le responsable de l'enquête parviendront sans doute à la conclusion que la méthode la plus pratique, la plus efficace et la plus économique est un échantillonnage en deux phases.

155. La post-stratification de l'échantillon primaire est importante pour deux raisons. La question ou les questions posées lors du quadrillage seront presque toujours succinctes car elles sont greffées sur une autre enquête qui comporte certainement déjà une longue entrevue. Le responsable de l'enquête principale n'acceptera sans doute pas qu'il soit ajouté en vue du quadrillage une série de questions très détaillées. Il est donc probable que certains des ménages, dans l'exemple susmentionné, identifiés comme comportant des orphelins n'en auront pas et inversement. Ces erreurs de classement portent à conclure qu'il faut prévoir deux strates, l'une pour les ménages ayant donné un résultat positif lors du quadrillage et l'autre pour les ménages ayant donné un résultat négatif. Des échantillons seraient alors prélevés dans chaque strate, l'idée étant qu'il y a probablement eu jusqu'à un certain point des erreurs de classement. Le taux d'échantillonnage dans la strate « oui » serait très élevé, jusqu'à 100 %, et un échantillon beaucoup plus réduit serait pris dans la strate « non ».

3.9.2. Échantillonnage visant à estimer un changement ou une tendance

156. Dans beaucoup de pays, les enquêtes sur les ménages sont conçues dans le double but d'estimer : *a*) des indicateurs *de référence* (leurs *niveaux*) à l'occasion de la première administration de l'enquête; et *b*) le *changement* de ces indicateurs lors de la deuxième administration et des adminis-

trations suivantes de l'enquête. Lorsque l'enquête est répétée plus d'une fois, les tendances des indicateurs sont également mesurées. Lorsque des enquêtes sont répétées, il se produit plusieurs effets sur la conception de l'échantillon qui n'existent pas lorsqu'il est réalisé une enquête transversale ponctuelle. Les questions qui doivent retenir l'attention sont notamment la fiabilité des estimations du changement et la combinaison appropriée à suivre s'agissant de savoir s'il convient d'utiliser les mêmes ménages ou des ménages différents d'une enquête à l'autre. Une question connexe concerne les effets de distorsion qui peuvent être introduits et le fait que, pour les ménages, être interrogés maintes fois peut être pesant.

157. Pour examiner l'aspect fiabilité, il faut également une démonstration mathématique. Nous commencerons par analyser la variance du changement estimatif, $d = p_1 - p_2$ exprimée comme suit :

$$\sigma_d^2 = \sigma_{p1}^2 + \sigma_{p2}^2 - 2\sigma_{p1,p2} = \sigma_{p1}^2 + \sigma_{p2}^2 - 2\rho\sigma_{p1}\sigma_{p2} \qquad (3.22)$$

où p est la proportion devant être estimée; σ_d^2 est la variance de la différence; σ_p^2 est la variance de p à la première ou deuxième occasion, dénotée par 1 ou 2; $\sigma_{p1,p2}$ est la covariance entre p_1 et p_2; et ρ est la corrélation entre les valeurs observées de p_1 et de p_2 lors des deux occasions.

Dans tous les cas où le changement estimatif est relativement réduit, ce qui est souvent le cas, nous avons :

$$\sigma_{p1}^2 \approx \sigma_{p2}^2$$

Ainsi, $\sigma_d^2 = 2\sigma_p^2 - 2\rho\sigma_p^2$ (l'on peut supprimer les exposants 1 et 2). Donc :

$$\sigma_d^2 = 2\sigma_p^2 (1-\rho) \qquad (3.23)$$

158. Pour évaluer l'équation (3.22), il y a lieu de noter qu'une estimation de σ_p^2 pour une enquête en grappes est celle d'un échantillon aléatoire simple multiplié par l'effet de conception de l'échantillon, *deff*. La corrélation, ρ, qui est la plus forte lorsque l'on utilise un échantillon de ménages, peut être égale à 0,8, voire encore plus élevée. Dans ce cas, l'estimateur, s_d^2, de σ_d^2 est donné par l'équation :

$$s_d^2 = 2\left[(pq)f/n\right](0{,}2), \quad \text{ou} \quad 0{,}4\,(pq)f/n \qquad (3.24)$$

159. Si l'on utilise les mêmes grappes mais des ménages différents, ρ demeure positif mais est beaucoup plus petit, peut-être de l'ordre de 0,25 à 0,35. Nous aurions alors (pour ρ de 0,3) :

$$s_d^2 = 2\left[(pq)f/n\right](0{,}7), \quad \text{ou} \quad 1{,}4\,(pq)f/n \qquad (3.25)$$

160. Enfin, en utilisant un échantillon totalement indépendant à la deuxième occasion ainsi que des grappes différentes et des ménages différents, ρ est égal à zéro et nous avons :

$$s_d^2 = 2\left[(pq)f/n\right] \qquad (3.26)$$

En utilisant une valeur caractéristique pour le *deff*, à savoir 2,0, la formule 3.26 donne :

$$s_d^2 = 4\left[(pq)/n\right] \tag{3.27}$$

161. Dans le cas d'enquêtes répétées se chevauchant en partie, par exemple 50 % de grappes et de ménages déjà interrogés et 50 % de nouveaux ménages, ρ doit être multiplié par un facteur F égal à la proportion que représente le chevauchement. Ainsi, l'équation 3.23 devient :

$$\sigma_d^2 = 2\sigma_p^2 (1 - F\rho) \tag{3.28}$$

162. Ce qui précède appelle quelques observations intéressantes. Premièrement, la variance estimée d'un changement estimatif relativement modeste entre deux enquêtes portant sur le même échantillon de ménages n'est que de 40 % environ de la variance du niveau à la première ou à la deuxième occasion. Si l'on utilise les mêmes grappes mais des ménages différents, l'on obtient une variance estimative du changement qui est de 40 % *supérieure* à celle du niveau. Des échantillons indépendants produisent une variance estimative représentant le *double* de celle du niveau.

163. Utiliser les mêmes ménages lors d'enquêtes répétées a donc beaucoup d'avantages sur le plan de la fiabilité. Si ce n'est pas possible, l'on peut néanmoins obtenir de bien meilleurs résultats en utilisant soit : *a*) une partie des mêmes ménages; soit *b*) les mêmes grappes mais avec des ménages différents. Les deux stratégies produisent des estimations caractérisées par une variance plus faible que celle de la formule la moins attrayante, consistant à utiliser des échantillons totalement indépendants.

164. S'agissant de la question des erreurs autres que d'échantillonnage, plus on répète l'utilisation du même échantillon de ménages, et plus fréquents sont deux effets négatifs dus aux déclarants, à savoir des non-réponses et des réponses conditionnées. Non seulement les déclarants répugnent-ils de plus en plus à coopérer, ce qui accroît les taux de non-réponses à mesure que les enquêtes se répètent, mais encore ils se trouvent peu à peu conditionnés, de sorte que la qualité et l'exactitude de leurs réponses peut se dégrader à mesure que les entrevues se multiplient.

165. Un phénomène connexe est celui de la « distorsion liée aux dates d'entrevue », qui a pour effet que l'on obtient des estimations différentes des déclarants qui fournissent des réponses portant sur la même période, mais qui ne participent pas à l'enquête depuis la même date. Ce phénomène a fait l'objet d'études approfondies et se retrouve dans les enquêtes portant sur des sujets très divers comme la population active, les dépenses, les recettes ou la victimisation. Aux États-Unis d'Amérique, par exemple, où les déclarants, dans les enquêtes sur la population active, sont interrogés huit fois, les estimations du chômage, dans le cas des déclarants interrogés pour la première fois, sont toujours supérieures de 7 % environ aux chiffres moyens obtenus des déclarants sur l'ensemble des huit entrevues. Ce phénomène a été constaté aux États-Unis pendant de nombreuses années de suite. Pour expliquer cette distorsion, les experts ont suggéré, entre autres, ce qui suit :

- Il se peut que les enquêteurs ne stimulent pas autant les déclarants lors des entrevues qui suivent la première.
- Il se peut que les déclarants, s'étant rendu compte que certaines réponses entraînent des questions supplémentaires, évitent de fournir certaines réponses.

- La première entrevue peut porter sur des événements n'entrant pas dans la période de référence, tandis que, dans le cas des entrevues suivantes, l'événement est « délimité dans le temps ».
- L'enquête peut en fait conduire les déclarants à changer de comportement.
- Il peut arriver que les déclarants ne fassent pas le même effort pour fournir les réponses exactes, lorsque le processus d'enquête commence à les lasser (Kasprzyk, 1989).

166. Il y a lieu de noter que la plupart des raisons indiquées ci-dessus visent des entrevues répétées menées pour la même enquête mais que, lorsque les mêmes ménages sont utilisés pour des enquêtes différentes, certains des traits de comportement relevés parmi les déclarants apparaissent aussi.

167. Il découle de ce qui précède qu'il faut prévoir des effets concurrents lorsque l'on utilise :

a) Le même échantillon de ménages à chaque occasion;

b) Des ménages de remplacement pour une partie de l'échantillon;

c) Un nouvel échantillon de ménages chaque fois que l'enquête est administrée.

168. À mesure que l'on avance de *a* à *c,* l'erreur d'échantillonnage qui caractérise les estimations du changement augmente tandis que les erreurs autres que d'échantillonnage tendent à diminuer. L'erreur d'échantillonnage est la plus faible lorsque l'on utilise le même échantillon de ménages à chaque occasion étant donné que la corrélation entre les observations est alors la plus forte. En revanche, l'utilisation des mêmes ménages aggrave la distorsion autre que celles dues à l'échantillonnage. Ce serait l'inverse lorsque l'on utilise à chaque occasion un nouvel échantillon de ménages.

169. C'est l'option *b* qui est généralement considérée comme celle qui offre le meilleur compromis s'agissant de concilier les erreurs d'échantillonnage et les distorsions non imputables à l'échantillonnage. Si l'on conserve une partie de l'échantillon année après année, l'erreur d'échantillonnage est moindre que dans le cas *c* et les erreurs autres que d'échantillonnage sont moindres que dans le cas *a*. Lorsqu'une enquête est menée à deux occasions seulement, l'option à privilégier est sans doute l'option *a*. Les effets imputables aux déclarants, en effet, n'auront sans doute pas d'impact trop marqué sur l'erreur totale lorsque l'échantillon n'est utilisé qu'à deux reprises. En revanche, si l'enquête doit être répétée trois fois ou davantage, il vaut mieux retenir l'option *b*. Une stratégie commune consiste à remplacer 50 % de l'échantillon à chaque occasion, par roulement (voir le chapitre 4 pour des exemples d'échantillonnage par roulement dans le contexte d'échantillons-maîtres).

3.10. Incidents d'exécution

170. L'on trouvera dans la présente section un résumé des mesures à adopter lorsque la mise en œuvre du plan d'échantillonnage se heurte à des obstacles, dont la plupart ont déjà été analysés et évoqués plus haut. Cependant, l'un des principes les plus importants de tous ceux qui sont soulignés dans le présent chapitre et dans le chapitre suivant est que nombre des obstacles peuvent être évités si l'enquête est planifiée très soigneusement au stade de la conception de l'échantillon. Des problèmes imprévus peuvent néanmoins surgir même avec la meilleure planification.

3.10.1. Définition et couverture de la population cible

171. Des problèmes se posent souvent, pour des raisons très diverses, lorsque la population effectivement couverte par l'enquête n'est pas la population cible visée.

Exemple

Prenons le cas d'une enquête devant couvrir la population cible type, à savoir tous les habitants du pays. La population effectivement couverte (c'est-à-dire la population parmi laquelle l'échantillon est sélectionné) est souvent inférieure à la population totale, pour l'une quelconque des raisons ci-après :

- Les personnes vivant dans des établissements comme hôpitaux, prisons et casernes ne sont pas prises en compte.
- Les personnes résidant dans certaines régions géographiques peuvent être délibérément exclues. Il peut s'agir de zones difficiles d'accès, de zones affectées par des catastrophes naturelles, de régions dont l'accès est interdit en raison de troubles civils ou d'hostilités, de complexes ou de camps où vivent des réfugiés ou des travailleurs étrangers, etc.
- Les personnes qui n'ont pas de résidence permanente sont considérées comme sortant du champ de l'enquête; tel peut notamment être le cas de populations nomades, de bateliers, de travailleurs itinérants, etc.

172. Le problème que soulèvent ces sous-populations dans le contexte du plan d'échantillonnage est qu'elles ne sont généralement pas identifiées d'emblée comme des groupes devant être exclus. Des difficultés surgissent par conséquent au stade de l'exécution lorsque l'échantillon sélectionné englobe par inadvertance, par exemple : *a*) une grappe qui s'avère être un camp, une prison ou un dortoir, plutôt qu'une zone résidentielle « classique »; ou *b*) une UPE qui se trouve dans un secteur montagneux et qui est considérée comme inaccessible. La « solution » fréquemment adoptée en pareille situation consiste à prendre une autre UPE. Cette solution, toutefois, entraîne une distorsion.

173. La meilleure solution consiste à éviter le problème au stade de la conception de l'échantillon, essentiellement en définissant soigneusement la population cible et en spécifiant non seulement quelles sont les sous-populations que celle-ci comprend mais aussi celles qui doivent être exclues. Deuxièmement, le calendrier d'échantillonnage doit alors être modifié pour en extraire les zones géographiques qui ne doivent pas être couvertes par l'enquête. Cela vaut également pour toute zone d'énumération particulière, par exemple un camp de travail, à exclure. Troisièmement, l'échantillon doit être sélectionné à partir du cadre ainsi modifié. L'on trouvera au chapitre 4 une analyse plus détaillée des cadres d'échantillonnage.

174. Il ne faut pas perdre de vue non plus que la solution suggérée plus haut permet de définir la population cible avec plus de précision. Il importe de décrire la population cible avec exactitude dans les rapports d'enquête de manière que l'usager en soit conscient.

3.10.2. Échantillons trop nombreux pour le budget disponible

175. Un autre problème surgit lorsque la taille de l'échantillon est plus grande que ce que permet le budget. En pareil cas, les responsables de l'enquête doivent ou bien obtenir des crédits supplémentaires, ou bien modifier les objectifs de la mesure en réduisant soit la précision, soit le nombre de domaines.

176. Un moyen de réduire la précision (accroître l'erreur d'échantillonnage) pour réduire beaucoup le coût de l'enquête consiste à sélectionner moins d'UPE tout en conservant un échantillon de même taille. Par exemple, plutôt que 600 UPE de 15 ménages chacune (n = 9 000), le plan d'échantillonnage pourrait être modifié de façon à sélectionner 400 UPE de 22 ou 23 ménages chacune

($n \approx 9\ 000$). S'agissant des domaines, une solution peut consister à se contenter de quatre grandes régions du pays plutôt que, par exemple, de prendre dix provinces.

3.10.3. Taille de la grappe plus petite ou plus grande que prévu

177. Un problème qui se pose souvent est qu'une grappe peut être bien plus grande que la taille qu'elle devrait avoir, par exemple à la suite de la construction de nouveaux logements, surtout si le cadre d'échantillonnage est ancien. Il se peut que les enquêteurs s'attendent à trouver 125 ménages dans une grappe donnée mais constatent, lors de l'établissement de la liste, que celle-ci en compte en fait 400. Une solution plausible, en pareil cas, consiste à subdiviser la grappe en sous-segments géographiques de tailles à près égales en termes de population. Le nombre de segments devrait être égal au nombre actuel de ménages divisé par la taille initialement prévue, arrondie au nombre entier le plus proche. Dans notre exemple, cela serait 400/125 ou 3,2, arrondi à 3 segments. Les segments seraient créés par l'établissement d'une carte et d'un dénombrement rapide des habitations (par opposition aux ménages). L'on sélectionnerait ensuite un segment au hasard pour l'établissement d'une liste.

178. Le problème inverse peut également se poser. Il se peut qu'une grappe soit beaucoup plus petite que prévu par suite de la démolition de logements, d'une catastrophe naturelle ou d'autres raisons. L'on est alors souvent tenté de la remplacer par une autre grappe, mais cela a pour effet d'introduire une distorsion. Il faut plutôt conserver telle quelle la grappe plus petite. Cela peut certes déboucher sur un échantillon dont la taille ultime est plus réduite que celle visée, mais l'augmentation de l'erreur d'échantillonnage sera mineure à moins qu'il n'y ait un grand nombre de grappes dans cette situation. Retenir telle quelle la grappe plus petite sans modification (ou substitution) permettra néanmoins d'obtenir une estimation dépourvue de distorsion car la grappe « représente » le changement intervenu dans la population depuis l'élaboration du cadre.

3.10.4. Cas de non-réponse

179. Bien que ce problème se rapporte davantage à l'exécution de l'enquête qu'à la sélection de l'échantillon, les non-réponses constituent un problème sérieux qui peut gravement compromettre les estimations provenant de l'enquête sur les ménages (voir les chapitres 6 et 8 pour un examen détaillé du problème posé par les non-réponses). Si l'on se trouve en présence de non-réponses dans plus de 10 à 15 % des cas, il en résultera dans les estimations des distorsions qui pourront les remettre sérieusement en question. Dans ce cas également, nombre de pays ont tendance à « résoudre » le problème des non-réponses en remplaçant les non-déclarants par des ménages qui répondent effectivement. Cette méthode elle-même introduit une distorsion car les nouveaux ménages continuent de ne représenter que des déclarants, et pas les ménages qui n'ont pas répondu. Or, l'on sait que les caractéristiques de ces deux derniers groupes sont différentes pour ce qui est d'importantes variables, en particulier celles qui ont trait à la situation socioéconomique. La solution à préférer, qui, regrettablement, n'est jamais couronnée de succès à 100 %, consiste à obtenir des réponses des ménages qui, initialement, n'en ont pas fournies. Pour cela, il faut d'emblée prévoir de se mettre en rapport aussi souvent que nécessaire avec les ménages non déclarants pour essayer d'obtenir leur coopération (s'ils ont refusé de répondre) ou pour les trouver chez eux (dans le cas des absents ou des ménages non disponibles pour d'autres raisons). Il pourra y avoir jusqu'à cinq rappels, mais le minimum doit être de trois.

3.11. Résumé des lignes directrices à suivre

180. La présente section résume les principales lignes directrices à suivre. Certaines d'entre elles valent dans presque toutes les circonstances (par exemple « utiliser un échantillonnage probabiliste »), d'autres devront donner lieu à des exceptions, selon les circonstances, les ressources et les exigences du pays. Les lignes directrices sont donc présentées essentiellement comme un guide plutôt que comme des recommandations à appliquer dans tous les cas. Les personnes appelées à participer à l'enquête devront :

- Utiliser des méthodes d'échantillonnage probabiliste à toutes les phases de la sélection.
- S'employer, autant que possible, à concevoir un échantillon simple plutôt que complexe.
- Utiliser des techniques de sélection qui donnent des échantillons autopondérés ou à peu près autopondérés à l'intérieur des divers domaines, ou globalement si la conception ne comporte pas de domaines.
- Utiliser une conception en deux phases si possible
- Calculer la taille de l'échantillon en utilisant une formule comme l'équation (3.5) et en ajustant la valeur des paramètres fixes (comme le taux prévisible de non-réponses et la taille moyenne des ménages) selon que de besoin eu égard à la situation du pays.
- Utiliser pour l'effet de conception une valeur par défaut de 2,0, à moins que des informations plus précises soient disponibles.
- Fonder la taille de l'échantillon sur l'estimation clé dont on pense qu'elle englobe le pourcentage le plus faible de la population de toutes les estimations clés que devra établir l'enquête.
- Si le budget le permet, choisir pour l'estimation clé une marge d'erreur ou un degré de précision représentant 10 % de l'estimation, c'est-à-dire une erreur relative de 10 %, au *niveau de confiance de 95 %;* si cela n'est pas possible, se contenter d'une erreur relative de 12 à 15 %.
- Définir les unités primaires d'échantillonnage (UPE) comme étant les zones d'énumération (ZE) du recensement, si cela est commode et approprié.
- Utiliser une stratification implicite jointe à un échantillonnage systématique sur la base d'une *probabilité proportionnelle à la taille* dans tous les cas où cela est possible, surtout pour les enquêtes à fins multiples.
- Limiter le nombre de domaines d'estimation au minimum absolument nécessaire (de manière à ramener la taille de l'échantillon à un niveau gérable).
- Essayer d'obtenir le plus grand nombre possible, plusieurs centaines, de grappes (ou d'UPE dans le cas d'une sélection en deux phases).
- Utiliser des grappes de petites tailles (10 à 15 ménages); plus les grappes seront petites, et mieux cela vaudra.
- Utiliser pour les grappes une taille constante plutôt que variable, c'est-à-dire un nombre fixe plutôt qu'une proportion fixe de ménages.
- Pour les domaines, tendre à un minimum de 50 UPE chacun.
- Prévoir au minimum trois, mais de préférence cinq, rappels auprès des ménages non déclarants.

- Dans le cas des populations rares, envisager une approche d'échantillonnage en deux phases consistant à greffer une question sur une autre enquête déjà prévue de manière à localiser les personnes faisant partie du groupe cible, et prévoir ensuite une entrevue intensive avec les membres d'un sous-échantillon.
- Pour les enquêtes visant à mesurer un changement, n'interroger les mêmes ménages à deux occasions que s'il n'est prévu que deux entrevues; s'il en est prévu trois ou plus, utiliser un système de chevauchement partiel en remplaçant les anciens ménages par de nouveaux ménages, par roulement, à chaque occasion.

Références et autres lectures

Banque mondiale (1999). *Core Welfare Indicators Questionnaire (CWIQ) Handbook.* Washington, Banque mondiale, chapitre 4.

Bennett, S. (1993). The EPI cluster sampling method: a critical appraisal. Invited paper, International Statistical Institute Session, Florence (Italie).

Cochran, W. (1977). *Sampling Techniques,* troisième édition, New York, Wiley.

Fonds des Nations Unies pour l'enfance (2000). *End-Decade Multiple Indicator Survey Manual.* Chapitre 4, intitulé « Designing and selecting the sample » et appendice 7, intitulé « *Sampling details.* » New York, UNICEF.

Hansen, M., W. Hurwitz et W. Madow (1953). *Sample Survey Methods and Theory,* New York, Wiley.

Hussmans, R., F. Mehran et V. Verma (1990). *Surveys of Economically Active Population, Employment, Unemployment and Underemployment: An ILO Manual on Concepts and Methods*. Chapitre 11 « Sample design », Bureau international du Travail, Genève.

International Statistical Institute (1975). *Manual on Sample Design*. World Fertility Survey Basic Documentation. Voorburg (Pays-Bas).

Kalton, G. (1983). *Introduction to Survey Sampling.* Beverly Hills, Californie, Sage. Publications.

_____ (1987). An assessment of the WHO Simplified Cluster Sampling Method for estimating immunization coverage. Rapport à l'UNICEF, New York.

_____ (1993). *Programme de mise en place de dispositifs nationaux d'enquêtes sur les ménages : Sampling Rare and Elusive Populations.* Département de l'information économique et sociale et de l'analyse des politiques et Division de statistique. INT-92-P80-16E. New York, Organisation des Nations Unies.

Kasprzyk, D. *et al.,* eds. (1989). *Panel Surveys.* New York, John Wiley & Sons, chap. 1.

Kish, L. (1965). *Survey Sampling,* New York, Wiley.

Krewski, D., R. Platek et J. N. K. Rao, eds. (1981). *Current Topics in Survey Sampling.* New York, Academic Press.

Le, T. et V. Verma (1997). *An Analysis of Sample Designs and Sampling Errors of the Demographic and Health Surveys,* DHS Analytical Reports, n° 3. Calverton, Maryland, Macro International Inc.

Ligue des États arabes (1990). *Sampling Manual, Arab Maternal and Child Health Survey,* Basic Documentation 5. Le Caire, Projet panarabe pour le développement de l'enfant (PAPCHILD).

Macro International Inc. (1996). *Sampling Manual.* DHS-III Basic Documentation, n° 6. Calverton, Maryland, Macro International Inc.

Namboodiri, N., ed. (1978). *Survey Sampling and Measurement.* New York, Academic Press.

Organisation des Nations Unies (1984). *Manuel d'enquêtes sur les ménages,* édition révisée, Études méthodologiques, n° 31. Numéro de vente : F.83.XVII.13.

_____ (1986). *Programmes de mise en place de dispositifs nationaux d'enquêtes sur les ménages : Sampling frames and sample designs for integrated household survey programmes.* New York, Organisation des Nations Unies, Département de la coopération technique pour le développement et Bureau de statistique.

_____ (2005). *Enquêtes sur les ménages dans les pays en développement et les pays en transition*. Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Organisation mondiale de la Santé (1991). *Expanded Programme on Immunization, Training for Mid-level Managers: Coverage Survey.* WHO/EPI/MLM91.10. Genève.

Raj, D. (1972). *Design of Sample Surveys.* New York, McGraw-Hill.

Scott, C. (1993), Discussant comments for session on « Inexpensive Survey Methods for Developing Countries ». Document établi en vue de la session de l'International Statistical Institute Session, Florence (Italie).

Som, R. (1966). *Practical Sampling Techniques*, deuxième édition. New York, Marcel Dekker, Inc.

Turner, A., R. Magnani et M. Shuaib, (1996). A not quite as quick but much cleaner alternative to the Expanded Programme on Immunization (EPI) Cluster Survey design. *International Journal of Epidemiology,* Liverpool (Royaume-Uni), vol. 25, n° 1.

United States Bureau of the Census (1978). *Current Population Survey Design and Methodology.* Technical Paper 40. Washington, Bureau of the Census.

Verma, V. (1991). *Sampling Methods*. Training Handbook. Tokyo, Statistical Institute for Asia and the Pacific.

Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association,* vol. 73, p. 40-46.

Chapitre 4

Cadres d'échantillonnage et échantillons-maîtres

4.1. Les cadres d'échantillonnage dans les enquêtes sur les ménages

1. Le chapitre précédent a traité, à l'exception des cadres d'échantillonnage, des multiples aspects de la conception des échantillons et de certaines des options pouvant être envisagées pour la conception des échantillons à utiliser pour une enquête sur les ménages. Cependant, l'un des aspects les plus importants de la conception d'un échantillon est le cadre d'échantillonnage, de sorte que ce sujet fait l'objet d'un chapitre distinct.

2. Le cadre d'échantillonnage a d'importantes incidences sur le coût et la qualité de l'enquête, qu'il s'agisse d'une enquête sur les ménages ou d'un autre type d'enquêtes. Dans les enquêtes sur les ménages, les défaillances du cadre d'échantillonnage sont l'une des plus fréquentes sources d'*erreurs autres que d'échantillonnage,* et en particulier d'erreurs dues à une sous-estimation d'importants sous-groupes de population. L'on essaiera, dans le présent chapitre, de définir les pratiques optimales à suivre en matière d'élaboration et d'usage des cadres, compte tenu des différentes phases de l'échantillonnage. Le chapitre est divisé en deux sections : la première est consacrée aux aspects généraux des cadres d'échantillonnage et de leur élaboration, l'accent étant mis sur la conception d'un échantillon en plusieurs phases pour une enquête sur les ménages, et la seconde évoque les questions particulières à prendre en considération lorsque l'on utilise un cadre *directeur* d'échantillonnage.

4.1.1. Définition[1] du cadre d'échantillonnage

3. Une définition opérationnelle simple d'un cadre d'échantillonnage est celle-ci : la *série de sources à partir desquelles l'échantillon est sélectionné.* La définition englobe également le but du cadre d'échantillonnage, qui est d'offrir un moyen de choisir les membres spécifiques de la population cible qui doivent être interrogés lors de l'enquête. Il peut être nécessaire d'avoir recours à plus d'une série de sources. Tel est généralement le cas des enquêtes sur les ménages en raison des multiples phases qu'elles comportent. Aux premières phases de la sélection, les échantillons sont habituellement prélevés de cadres constitués par des *zones* géographiques. Lors de la dernière phase, des échantillons peuvent être sélectionnés au moyen d'un cadre constitué d'une zone ou d'une *liste* (voir ci-après pour une discussion de ces deux types de cadres).

1 Le lecteur est invité à se référer au tableau 3.1 du chapitre 3 (p. 31) pour le glossaire des termes employés dans les chapitres 3 et 4.

4.1.1.1. Cadre d'échantillonnage et population cible

4. Une considération importante, pour décider du ou des cadres à utiliser aux fins d'une enquête sur les ménages, est la corrélation entre la population cible et l'unité de sélection. C'est en effet l'unité de sélection qui détermine le cadre, de même que la probabilité de sélection à la dernière phase.

> **Exemple**
>
> Dans le cas d'une enquête dont la population cible est les enfants en bas âge, l'on peut envisager deux cadres possibles : l'un serait les établissements médicaux ayant enregistré des naissances au cours des 12 mois écoulés, et l'autre les ménages comportant des enfants de moins de 12 mois. Dans le premier cas, le cadre comprendrait deux parties, une pour chaque phase de la sélection : premièrement, la liste des hôpitaux et des cliniques où sont réalisés des accouchements; et, deuxièmement, la liste de tous les enfants nés dans ces établissements au cours des 12 mois écoulés. Les unités de sélection seraient les établissements médicaux à la première phase et les nouveau-nés à la seconde. Ainsi, l'unité de sélection et la population cible sont des termes synonymes à la *dernière* phase de la sélection. Dans le second cas, cependant, le cadre serait probablement défini à une phase ultérieure de la sélection comme étant une liste de ménages de zones peu étendues comme des villages ou des pâtés de maisons. Lors de l'application du plan d'échantillonnage, il serait sélectionné des ménages déterminés comme ayant des enfants de moins de 12 mois. Dans ce cas, l'unité de sélection sur laquelle est fondée la probabilité de sélection serait le ménage. Il y a lieu de noter toutefois que les membres de la population cible ne sont pas effectivement identifiés et interrogés tant qu'il n'a pas été déterminé que les ménages comportent effectivement des nouveau-nés. Ainsi, dans le cas du cadre fondé sur les ménages, l'unité de sélection et la population cible sont différentes.

5. Dans le cas des enquêtes sur les ménages, qui sont l'objet du présent guide, l'unité de sélection ainsi que l'unité autour de laquelle est structurée la conception de l'échantillon est le ménage. Néanmoins, la population cible, même dans le cas d'une enquête de caractère général, sera différente, selon les objectifs de la mesure. Hormis les enquêtes sur les revenus et les dépenses des ménages, la population cible sera habituellement une population autre que celle du ménage lui-même. L'on peut en citer comme exemples les enquêtes sur l'emploi, dans le cas desquelles la population cible comprend généralement les personnes de 10 (ou 14) ans ou plus, ce qui exclut totalement les jeunes enfants, ou les enquêtes sur la santé génésique des femmes, dans le cas desquelles la population cible comprend les femmes de 14 à 49 ans (et souvent uniquement les femmes de ce groupe d'âge qui ont été mariées au moins une fois), etc.

4.1.2. Propriétés des cadres d'échantillonnage

6. Comme on l'a dit, il va de soi que les cadres d'échantillonnage doivent capturer, au sens statistique, la population cible. En outre, un cadre d'échantillonnage parfait est un cadre *complet, exact et à jour*. Ce sont là des propriétés idéales qui ne se retrouvent jamais dans les enquêtes sur les ménages. Il est néanmoins essentiel de s'en rapprocher autant que possible soit en élaborant un cadre à partir de zéro, soit en utilisant un cadre existant. La qualité du cadre peut être évaluée en déterminant dans quelle mesure ses propriétés idéales se rapportent à la population cible. Comme on l'a dit au chapitre 3, notre définition d'un échantillon probabiliste, c'est-à-dire d'un échantillon dans lequel chacun des membres de la population cible a une chance connue, autre que zéro, d'être sélectionné, est un paramètre utile pour juger de la qualité d'un cadre d'échantillonnage.

7. Selon l'impossibilité de parvenir à chacune des propriétés idéales, les résultats de l'enquête souffriront de différentes distorsions, allant toutefois souvent dans le sens d'une *sous-estimation* de la population cible.

4.1.2.1. Complétude

8. Le cadre idéal serait considéré comme complet, en ce qui concerne la population cible, si tous ses membres (c'est-à-dire l'*univers*) étaient couverts. La couverture de la population ou des populations cibles est par conséquent une caractéristique essentielle pour juger si le cadre est approprié. S'il ne l'est pas, donc, l'équipe chargée de l'enquête doit voir s'il peut être réparé ou développé pour qu'il soit plus complet. Dans l'exemple précédent, les enfants nés à la maison ou ailleurs que dans un établissement médical ne seraient pas couverts par l'enquête si les établissements médicaux constituaient le seul cadre d'échantillonnage. Dans cette exemple, il y aurait une forte proportion de la population cible qui aurait une chance zéro d'inclusion, et la condition nécessaire à un échantillon probabiliste serait absente. De ce fait, le cadre reposant sur les établissements médicaux continuerait à sous-estimer le nombre de nouveau-nés. De plus, les *caractéristiques* des nouveau-nés seraient sans doute fort différentes de celles des enfants nés à la maison. Le cadre fondé sur les établissements médicaux donnerait par conséquent des répartitions biaisées pour d'importants indicateurs concernant les nouveau-nés ou les soins dont ils font l'objet.

9. Une couverture insuffisante peut également constituer un problème dans le cas des enquêtes sur les ménages. Par exemple, un plan national peut être conçu de manière à englober l'ensemble de la population, mais il y a divers sous-groupes, comme les personnes qui vivent en établissement, les ménages nomades et les bateliers, qui ne vivent pas dans des ménages. En pareil cas, il n'est évidemment pas possible de couvrir l'ensemble de la population au moyen de l'enquête sur les ménages. Il faudra élaborer d'autres cadres pour couvrir les groupes autres que les ménages afin de donner à leurs membres une probabilité autre que zéro d'être inclus, faute de quoi la population cible effective devra être modifiée de manière à définir avec plus de précision les éléments qu'elle comprend. L'usager sera ainsi clairement informé des segments de la population qui ont été exclus de l'enquête.

4.1.2.2. Exactitude

10. L'exactitude est une autre caractéristique importante des cadres d'échantillonnage, bien que certains types de cadres risquent plus d'être inexacts que d'autres. L'on peut dire qu'un cadre est exact si chaque membre de la population cible est inclus une fois et une seule fois. Prenons l'exemple d'une liste des entreprises employant plus de 50 salariés. Il peut y avoir des erreurs si : *a*) une entreprise de la liste a 49 salariés ou moins; *b*) une entreprise qui a plus de 50 salariés manque; ou *c*) une entreprise figure sur la liste plus d'une fois (peut-être sous des noms différents).

11. Dans les enquêtes sur les ménages, de telles inexactitudes sont plus rares. Il peut néanmoins y en avoir, par exemple : *a*) si certains éléments sont absents d'un cadre composé d'un fichier informatique de zones d'énumération (ZE); *b*) si le cadre constitué par une liste des ménages d'un village a laissé de côté certains ménages vivant au périmètre du village; *c*) si, dans une liste de ménages de la zone d'énumération, certains des ménages figurent dans plus d'une unité; ou *d*) si une liste de ménages ancienne n'englobe pas les logements nouvellement construits. Ce dernier cas, celui d'un cadre qui n'est pas à jour, est discuté plus en détail ci-après.

12. Si des ZE ou des ménages manquent dans une zone d'énumération, cela signifie évidemment que les ménages en question n'ont aucune probabilité d'être sélectionnés pour l'échantillon. Dans ce cas également, l'une des conditions requises pour obtenir un échantillon véritablement probabiliste se trouverait absente. De doubles inscriptions violent également un critère de probabilité, à moins qu'il en soit tenu compte de sorte que l'on puisse calculer les probabilités réelles de sélection. Regrettablement, il arrive fréquemment que ces omissions et doubles inscriptions ne soient pas découvertes. Il se peut par conséquent que le technicien chargé de l'échantillonnage ne se rende pas compte qu'il faut corriger le cadre avant de l'utiliser pour l'échantillonnage. D'un autre côté, un petit nombre d'omissions ou de doubles inscriptions n'entraînent généralement pas de biais appréciable, ni même notable, dans les estimations.

4.1.2.3. Actualité

13. Il va de soi qu'idéalement le cadre doit être à jour pour posséder les deux autres propriétés, c'est-à-dire la complétude et l'exactitude. Un cadre obsolète contient évidemment des erreurs et sera généralement incomplet, surtout s'il est utilisé pour des enquêtes sur les ménages. L'exemple parfait de cadres obsolètes est un recensement de population remontant à plusieurs années. Ce recensement ne reflétera pas comme il convient les logements qui ont été nouvellement construits ou détruits ni les migrations dues aux naissances ou aux décès. Du fait de ces carences, la règle selon laquelle, dans un échantillon probabiliste, chaque membre de la population cible doit avoir une chance de sélection connue n'est pas respectée.

Exemple

Supposons que le cadre se compose de ZE définies sur la base du dernier recensement, qui remonte à quatre ans et qui n'a pas été mis à jour. Supposons en outre que de nombreux peuplements spontanés soient apparus dans les faubourgs de la capitales dans des ZE qui, à l'époque du recensement, étaient inhabitées ou presque inhabitées. La conception de l'échantillon n'offrirait aux ménages vivant dans les ZE précédemment inhabitées aucune chance d'inclusion, ce qui serait contraire aux conditions d'un échantillon probabiliste. Dans les ZE qui étaient presque inhabitées, il se poserait un autre problème sérieux même si, à proprement parler, les ZE en question ne violaient pas les conditions que doit refléter un échantillon probabiliste. L'échantillon serait indubitablement sélectionné sur la base d'une *probabilité proportionnelle à la taille,* celle-ci étant la population ou les ménages dénombrés lors du recensement. Comme la population était très réduite lors du recensement, une ZE ayant connu une forte croissance n'aurait qu'une chance réduite d'être sélectionnée sur la base d'un échantillon fondé sur la *probabilité proportionnelle à la taille*. De ce fait, la variance de l'échantillonnage pourrait atteindre des niveaux inacceptables.

4.1.3. Cadres constitués par des zones géographiques

14. Dans la présente section et dans la suivante, il est question de deux catégories de cadres utilisés pour l'échantillonnage, qu'il s'agisse d'enquêtes sur les ménages ou d'autres types d'enquêtes. Il importe de noter que, dans une conception à plusieurs phases, le cadre utilisé pour chacune d'elles doit être considéré comme une composante distincte. Le cadre spécifique est différent à chaque phase. La conception de l'échantillon pour une enquête sur les ménages utilisera généralement à la fois un

cadre fondé sur des zones géographiques (dont il est question dans la présente section) aux premières phases et un cadre fondé sur une liste (objet de la section suivante) lors de la dernière phase.

15. Dans les enquêtes sur les ménages, un cadre d'échantillonnage fondé sur des zones géographiques comprend toutes les unités géographiques du pays, rangées dans l'ordre hiérarchique. Ces unités sont désignées par des appellations administratives qui varient d'un pays à l'autre mais qui, dans l'ordre descendant, comprennent habituellement des termes comme province, comté, district, commune, village et hameau (régions rurales) ou quartier, arrondissement ou pâté de maisons (régions urbaines). Dans le cas des recensements, les subdivisions administratives sont également décomposées en zones maîtresses et zones d'énumération ou ZE. Souvent, les ZE définies aux fins des recensements constituent les unités géographiques les plus petites définies et délimitées dans un pays.

16. Les unités géographiques comportent quatre caractéristiques distinctes qui sont importantes pour la conception de l'échantillon :

a) Elles englobent habituellement l'ensemble du territoire d'un pays;

b) Leurs frontières sont généralement bien délimitées;

c) L'on dispose d'informations concernant leur population;

d) Il en existe une carte.

17. Il importe que l'enquête couvre l'intégralité du territoire du pays, comme on l'a vu, car c'est une des conditions qui doit être remplie si l'on veut avoir un bon échantillon probabiliste. Des frontières bien délimitées figurant sur une carte sont inappréciables lors de l'échantillonnage car elles permettent de cerner où le travail sur le terrain doit être réalisé. Lorsqu'il dispose d'informations exactes sur les limites des zones géographiques, l'enquêteur peut également localiser plus facilement les ménages faisant partie de l'échantillon qui devront être interrogés. Les effectifs de la population doivent être connus lors de la conception de l'échantillon pour déterminer la taille et calculer les probabilités de sélection.

18. L'on commence généralement, pour élaborer le cadre géographique qui sera utilisé lors d'une enquête sur les ménages, par prendre comme point de départ les informations provenant du recensement de la population, sur la base des quatre caractéristiques susmentionnées. En outre, la ZE est une unité géographique de taille appropriée qui servira à procéder à la sélection lors des phases ultérieures de l'échantillonnage (l'avant-dernière phase dans le cas d'une conception en deux phases). Dans la plupart des pays, les ZE sont délibérément construites de manière à contenir un nombre de ménages à peu près égal, souvent une centaine, afin d'égaliser la charge de travail des enquêteurs chargés du recensement.

19. Un cadre géographique est aussi, paradoxalement, une *liste,* car il faut commencer par une liste des circonscriptions administratives où vit la population pour mener à bien les premières phases de la sélection d'un échantillon. Cela conduit à analyser de manière un peu plus approfondie les cadres constitués par des listes.

4.1.4. Cadres constitués par des listes

20. Un cadre d'échantillonnage constitué par une liste est simplement un cadre composé d'une liste des unités de population cibles. Théoriquement, il existe un tel cadre pour chaque pays immédiatement après un recensement. En principe, le dernier recensement donne une liste, décomposée par régions géographiques de tous les ménages ou logements du pays.

21. La liste provenant d'un recensement récent est idéale comme cadre d'échantillonnage car elle constitue la liste la plus à jour, la plus complète et la plus exacte que l'on puisse imaginer. Comme elle est structurée sur une base géographique, il est assez simple de stratifier la liste de manière à gérer une répartition géographique appropriée de l'échantillon. Par conséquent, lorsqu'il faut mener après un recensement une enquête par sondage tendant à obtenir des informations supplémentaires ou des informations plus détaillées que celles qui peuvent raisonnablement être obtenues au moyen d'un recensement, cette nouvelle liste constitue un cadre d'échantillonnage idéal. Il importe de ne pas perdre de vue, cependant, que cette nouvelle liste ne demeure un *cadre à jour* que brièvement. Manifestement, plus l'intervalle entre le recensement et l'enquête de suivi est long, et moins la liste devient utile comme source d'informations pour l'établissement du cadre d'échantillonnage.

22. Il y a d'autres listes qui, selon leur qualité, peuvent aussi être considérées comme un cadre d'échantillonnage approprié pour une enquête sur les ménages, comme les registres de l'état civil et les registres de la compagnie d'électricité. Les registres de l'état civil peuvent être utilisés comme cadres dans les pays qui tiennent des états détaillés sur les citoyens et leur adresse. Dans certains cas, ces registres peuvent être plus utiles que le cadre géographique provenant d'un recensement car, généralement, ils sont tenus continuellement à jour. Les registres des compagnies d'électricité peuvent être utiles comme cadre d'échantillonnage lorsque le recensement est ancien, mais ils doivent évidemment être évalués pour identifier les problèmes potentiels et leur impact. Un problème évident, qui se traduirait par une sous-estimation, serait celui des ménages non raccordés à l'électricité, et un autre serait d'un seul raccordement desservant plusieurs ménages.

23. Un autre cadre largement utilisé dans les pays développé est un répertoire des abonnés au téléphone. L'échantillonnage se fait par *appels aléatoires* pour veiller à ce que les abonnés dont le numéro n'est pas publié aient eux aussi une probabilité appropriée d'être sélectionnés. Cette méthode n'est cependant pas recommandée dans les pays où la pénétration des services téléphoniques est réduite.

24. Dans une enquête sur les ménages de type classique, la dernière phase de la sélection est invariablement fondée sur un cadre d'échantillonnage constitué par une liste. Nous avons déjà vu comment l'avant-dernière phase peut donner un échantillon de grappes en fonction desquelles est établie une liste des ménages. C'est à partir de cette liste que les ménages devant faire partie de l'échantillon sont sélectionnés. Nous avons ainsi un cadre géographique définissant les grappes et une liste définissant, à l'intérieur des grappes, les ménages faisant partie de l'échantillon.

4.1.5. Cadres multiples

25. L'on a vu au chapitre 3 comment on procède à un échantillonnage en deux phases pour une enquête sur les ménages. Il est utilisé des méthodes de quadrillage pour identifier un groupe cible déterminé pendant la première phase et il est prévu ensuite, lors de la seconde phase, d'interroger un sous-échantillon du groupe identifié. Une autre méthode d'échantillonnage qui peut parvenir au même résultat, pour l'essentiel, consiste à utiliser plus d'un cadre d'échantillonnage. Habituellement, il n'en est utilisé que deux, et l'on est alors en présence d'une conception à *double cadre;* parfois, cependant, il peut être utilisé trois cadres, voire davantage (conception à *cadres multiples*). Par exemple, un cadre de population qui est défini comme étant une somme de plusieurs listes sur la base desquelles sera sélectionné un échantillon indépendant, en l'occurrence chaque sous-cadre, devient une strate (pour une analyse de la stratification, voir la section 3.4.2 du chapitre 3 et l'annexe I). Le problème que soulèvent habituellement de tels cadres, cependant, est celui des doubles emplois.

4.1.5.1. Double cadre habituellement utilisé pour les enquêtes sur les ménages

26. Pour simplifier l'exposé, nous ne traiterons que des conceptions à double cadre, bien que les principes soient les mêmes pour les conceptions à cadre multiple. D'une manière générale, cette méthode consiste à combiner un cadre géographique englobant l'ensemble de la population et une liste de personnes dont on sait qu'elles font partie de la population cible à l'examen. Prenons par exemple le cas d'une enquête visant à étudier les caractéristiques des chômeurs. L'enquête peut être fondée sur le cadre géographique de ménages, mais celui-ci peut être complété par un autre cadre constitué par la liste des personnes actuellement au chômage inscrites auprès des services du Ministère du travail. L'objectif d'un échantillon à double cadre de ce type est de parvenir à un échantillon de la taille voulue comprenant des personnes qui ont une très forte probabilité d'appartenir à la population cible. Cette approche peut être moins chère et plus efficace qu'un échantillonnage en deux phases. Il faut pour cela utiliser le cadre général de ménages afin de ne pas laisser de côté les membres de la population cible qui ne figurent pas sur la liste. Dans cet exemple, il s'agirait des chômeurs qui ne sont pas inscrits auprès des services du Ministère du travail.

27. Les conceptions à double cadre supposent néanmoins plusieurs limitations, dont l'une est que la liste doit être virtuellement à jour. Si une forte proportion de personnes sélectionnées sur la liste ont changé de statut de sorte qu'elles ne font plus partie de la population cible, la liste n'est de guère d'utilité. Dans notre exemple, le fait qu'un chômeur ayant trouvé un emploi au moment de l'enquête ne doit pas faire partie de la population cible illustre pourquoi la liste doit être à jour.

28. Une autre limitation tient au fait que les lieux de résidence des personnes figurant sur la liste sont généralement dispersés dans l'ensemble de la collectivité, de sorte que les interroger coûte cher en raison des frais de déplacement. Tel n'est évidemment pas le cas avec un cadre géographique, l'échantillon pouvant être sélectionné en grappes de manière à réduire le coût des entrevues.

29. Un des problèmes les plus sérieux que soulèvent les conceptions à double cadre est celui des doubles emplois. D'une façon générale, les personnes qui figurent sur la liste font également partie du cadre géographique. Dans ce cas également, dans notre exemple, les chômeurs sélectionnés sur un registre sont membres de ménages. Ils ont par conséquent une double chance de sélection lorsque les deux cadres sont utilisés. L'on peut ajuster le cadre en conséquence, mais cela a des incidences sur le contenu du questionnaire. Dans notre exemple, chaque chômeur interviewé dans le cadre de l'échantillon de ménages devrait être interrogé sur le point de savoir s'il est inscrit sur le registre des chômeurs tenu par le Ministère du travail. Dans le cas de ceux qui répondent par l'affirmative, un travail supplémentaire s'impose : il faut retrouver leur nom sur la liste, processus qui non seulement est complexe mais encore peut entraîner des erreurs. Lorsque leur nom est retrouvé, il faut modifier la pondération de la personne intéressée pour qu'elle soit égale à $(1/ P_h + 1/ P_l)$ de manière à refléter le fait que cette personne a une probabilité P_h d'être sélectionnée à partir du cadre de ménages et une probabilité P_l d'être sélectionnée à partir du cadre constitué par la liste. Il importe de noter que le rapprochement doit être fait pour l'ensemble de la liste et pas seulement pour les personnes faisant partie du cadre qui se trouvent avoir été sélectionnées dans l'échantillon. En effet, la probabilité (et la pondération) sont fonction des chances de sélection, qu'il y ait ou non sélection dans la pratique.

4.1.5.2. Cadres multiples pour différents types de logements

30. L'on peut utiliser une autre méthode d'échantillonnage à double cadre lorsque la population cible réside dans divers types de logements qui ne se chevauchent pas. Par exemple, une enquête sur

les orphelins sera généralement conçue de manière à englober les orphelins qui vivent dans deux types de logements : le premier cadre serait les établissements, par exemple les orphelinats, et le second les ménages, afin de dénombrer les orphelins qui vivent avec un parent survivant, d'autres membres de la famille ou des personnes ne faisant pas partie de la famille. L'échantillon serait conçu sur la base d'un double cadre, à savoir un cadre de ménages et un cadre institutionnel, l'un et l'autre ne se chevauchant évidemment pas.

31. L'objectif d'une conception de ce type est de couvrir comme il convient la population cible (de manière à parvenir à un pourcentage aussi proche de 100 % que possible). Lorsqu'une proportion significative de la population vit dans l'un ou l'autre des deux types de logements, il peut y avoir de sérieuses distorsions si l'échantillon est sélectionné au moyen de l'un des deux cadres seulement. Par exemple, un échantillon d'orphelins fondé seulement sur ceux qui vivent dans des ménages non seulement aboutirait à une sous-estimation de la population d'orphelins mais encore à une estimation biaisée pour ce qui est de leurs caractéristiques. Des distorsions semblables se produiraient dans le cas d'une enquête sur les orphelins vivant en établissement.

32. La limitation susmentionnée concernant les doubles emplois n'est pas applicable aux conceptions fondées sur un double cadre sans chevauchement. Ce dernier type de conception est donc beaucoup plus facile à administrer.

4.1.6. Cadre(s) type(s) dans les conceptions en deux phases

33. Le chapitre 3 a insisté sur l'intérêt pratique des conceptions en deux phases. La présente section est consacrée au cadre habituellement utilisé pour ce type de conception.

34. Les unités géographiques ou grappes sélectionnées à la première phase de la sélection sont fréquemment définies comme étant des villages (ou des parties de villages) ou des ZE du recensement dans les régions rurales et des pâtés de maisons dans les régions urbaines. Le cadre comprend par conséquent les unités géographiques qui constituent l'univers à l'étude, quelle que soit la définition qui en est donnée : l'ensemble du pays, une province ou une série de provinces ou de chefs-lieux. L'échantillonnage est réalisé par une compilation de la liste d'unités, laquelle est vérifiée pour déterminer si elle est complète, une stratification de la liste selon les modalités appropriées (fréquemment sur une base géographique) puis une sélection (habituellement sur la base d'une probabilité proportionnelle à la taille) d'un échantillon systématique des unités.

35. Si l'ensemble de grappes sélectionnées dans l'univers est très nombreux, il pourra être nécessaire de prévoir des phases intermédiaires fictives de sélection, comme discuté dans le chapitre précédent. En pareil cas, les unités seront définies de façon différente pour chacune des phases fictives. Dans l'exemple précédent concernant le Bangladesh, les unités prises comme cadre pour les deux phases fictives étaient définies comme les thanas et les districts.

36. Dans une conception en deux phases, les unités constituant le cadre appliqué pour la seconde phase sont simplement les ménages faisant partie des groupes sélectionnés comme échantillon pendant la première phase. Lorsque l'échantillon est constitué à partir d'une liste de ménages, le cadre est, par définition, une liste. L'échantillon peut également être constitué sur la base de segments compacts créés par une subdivision des grappes en éléments géographiques exhaustifs qui s'excluent mutuellement. En pareil cas, le cadre utilisé pour la seconde phase est un cadre géographique.

4.1.7. Cadres directeurs d'échantillonnage

37. L'on se contentera ici de mentionner brièvement le concept de cadre directeur d'échantillonnage, qui est analysé beaucoup plus en détail dans la section 4.2 ci-après.

38. Un cadre directeur est un cadre utilisé pour sélectionner les échantillons destinés soit à des enquêtes multiples ayant chacune un contenu différent, soit à différentes séries d'une enquête continue ou périodique. Hormis une mise à jour périodique, selon que de besoin, l'échantillonnage lui-même ne varie pas d'une enquête à l'autre ou d'une série à la suivante de la même enquête. Au contraire, c'est là sa caractéristique distinctive, le cadre directeur est conçu et construit de manière à être un cadre établi et stable permettant de sélectionner les sous-échantillons requis pour des enquêtes déterminées ou plusieurs séries de la même enquête devant être réalisées sur une période assez longue.

4.1.8. Problèmes que soulèvent communément les cadres et remèdes suggérés

39. Les problèmes que soulève l'utilisation de cadres défectueux lors d'enquêtes sur les ménages tiennent à la fois aux distorsions autres que celles qui sont dues à l'échantillonnage et à la variance de l'échantillonnage. Comme on l'a déjà dit, les problèmes surgissent lorsque le cadre d'échantillonnage est obsolète, inexact ou incomplet. Dans la grande majorité des enquêtes de caractère général d'envergure nationale, le cadre de base est le dernier recensement de la population, et tel est le cadre visé dans la présente section. Les cadres issus d'un recensement soulèvent fréquemment des problèmes parce qu'ils sont obsolètes, inexacts et incomplets, et l'ampleur de ces problèmes tend à s'aggraver à mesure que l'intervalle entre le recensement et l'enquête s'allonge.

40. L'on a déjà dit qu'un cadre doit être à jour pour refléter la population actuelle. Un cadre fondé, par exemple, sur un recensement remontant à cinq ans ne reflète pas comme il convient l'accroissement démographique ni les migrations. Même un cadre à jour peut être incomplet et entraîner des problèmes s'il n'englobe pas les casernes, les bateliers, les nomades et d'autres importantes sous-populations qui ne vivent pas dans des logements traditionnels. Les inexactitudes que comportent aussi bien les cadres à jour que les cadres provenant de recensements anciens soulèvent différentes difficultés, en particulier celles qui tiennent à un double dénombrement des ménages, à des ménages manquants ou à des ménages énumérés ou codés selon la ZE erronée.

41. Les stratégies à suivre en présence de cadres de recensement anciens, inexacts ou incomplets dépendent en partie : *a*) des objectifs de l'enquête; et *b*) de l'âge du cadre. S'agissant des objectifs de la mesure, si une enquête est délibérément conçue de manière à ne couvrir, par exemple, que les ménages stationnaires, un cadre excluant les ménages nomades suffira. D'un autre côté, il faudra mettre au point une procédure afin de créer un cadre de ménages nomades si l'enquête doit les englober (dans des pays où de telles populations existent). De ce point de vue, la question de savoir si un cadre de recensement est complet ou non dépend de la définition de la population ou des sous-populations cibles devant être couvertes par l'enquête.

42. Les solutions à adopter pour remédier aux problèmes d'obsolescence et d'inexactitude varieront selon l'âge du recensement. S'il n'est peut-être pas prudent de proposer une règle précise étant donné la diversité des conditions qui prévalent dans les divers pays, une règle approximative pouvant guider la stratégie à retenir pour refondre ou mettre à jour le cadre consisterait à déterminer si le recensement remonte à plus de deux ans. S'agissant des inexactitudes mentionnées au paragraphe 40, l'on peut avoir recours aux solutions évoquées dans les sous-sections 4.1.2 et 4.1.8.

4.1.8.1. Cadres de recensement remontant à plus de deux ans

43. La première situation est celle des pays dont les recensements remontent à deux ans ou plus. Ce sont ces cadres anciens qui soulèvent le plus de difficultés lorsqu'il s'agit de concevoir un échantillon à utiliser dans une enquête sur les ménages, surtout dans les villes en expansion rapide. La solution idéale consiste à actualiser totalement ce cadre ancien pour l'ensemble du pays car, si l'on y parvient, les données seront à la fois aussi exactes, en termes de couverture, et aussi fiables que possible. Regrettablement, c'est aussi la formule la plus longue et la plus onéreuse, de sorte qu'elle est difficile à appliquer dans la pratique. Il n'en demeure pas moins que cela peut être la seule solution dans les pays où les données provenant du recensement sont sérieusement dépassées.

44. Plutôt que de procéder à une mise à jour complète, une solution de compromis peut consister à mettre à jour le cadre uniquement dans les zones ciblées, celles-ci étant identifiées par des experts nationaux connaissant bien les schémas de croissance et les changements démographiques. Il est assez simple de mettre à jour le cadre de recensement. Ce qu'il faut, c'est mesurer la taille actuelle. Aux fins de la mise à jour du cadre, la taille serait définie comme étant le nombre de logements, par opposition au nombre de ménages ou de personnes.

45. Il importe de reconnaître que la taille n'a pas à être précise pour que la méthode d'échantillonnage soit valable. Par exemple, si, sur la base du dernier recensement, l'on pensait qu'une zone d'énumération donnée comptait 122 ménages, il n'y a pas de raison de s'inquiéter si elle en comporte aujourd'hui 115 ou 132. Aussi n'est-il guère utile de vouloir mettre à jour le cadre dans les quartiers établis de longue date qui n'ont guère changé depuis des décennies, même si des habitants du quartier vont et viennent. L'attention doit porter plutôt sur toute différence marquée entre la situation actuelle et celle qui prévalait lors du dernier recensement, par exemple si l'on trouve 250 ménages alors que l'on s'attendait à n'en trouver que 100. De telles situations sont fréquentes dans les quartiers où beaucoup de logements sont construits ou démolis, par exemple les communautés spontanées des faubourgs, les quartiers de grands immeubles et les sites de démolition. C'est pour de telles zones qu'il y a lieu de procéder à une mise à jour. Il faudra s'en remettre aux collaborateurs et aux experts nationaux pour aider à identifier les zones en question et, comme il va de soi, mettre à jour le cadre uniquement pour les zones où il y a eu des changements depuis le recensement.

46. Cette mise à jour comporte généralement plusieurs étapes, dont : *a*) l'identification des zones d'énumération qui constituent des zones cibles; *b*) un quadrillage rapide des zones d'énumération affectées afin d'en déterminer la taille actuelle; et *c*) la révision des fichiers du recensement pour y refléter la taille nouvellement identifiée. Le fait que, comme on l'a vu, une taille approximative suffit est la raison pour laquelle l'opération de quadrillage doit être menée de manière à identifier des logements plutôt que des ménages. Pour cela, il n'est pas nécessaire de frapper à la porte pour dénombrer des logements, sauf peut-être dans le cas d'immeubles où il faut entrer pour pouvoir compter le nombre de logements.

47. Pour mettre à jour de la sorte des cadres anciens, il faut stabiliser les probabilités de sélection à l'avant-dernière phase et par conséquent la fiabilité des estimations. Dans la pratique, l'opération de mise à jour aide à contrôler non seulement la taille et l'échantillon global mais aussi le travail que l'établissement de listes et les entrevues représenteront pour le personnel de terrain. De plus, l'on peut ainsi réduire le risque de découvrir sur le terrain des grappes qui s'avèrent être beaucoup plus grandes que prévu et de devoir en tirer des sous-échantillons. À ce propos, le fait que le sous-échantillonnage exige des ajustements de pondération complique le traitement des données. Or, ce risque

peut être réduit à tel point qu'il n'est pas trouvé de grappes de taille inattendue après la sélection de l'échantillon à l'avant-dernière phase.

48. Lorsqu'il faut concevoir un échantillon pour une enquête sur les ménages, il faut généralement établir une liste à jour des ménages qui comportent les grappes faisant partie de l'échantillon. Ainsi, il faut fréquemment une mise à jour à l'avant-dernière phase de la sélection (voir dans le point 4.1.8.2 ci-après la référence à l'échantillonnage fondé sur une probabilité proportionnelle à la taille estimative, considérations qui valent ici aussi). Par conséquent, la dernière liste des grappes qui n'ont pas été mises à jour peut être très proche des listes du recensement (bien que cela ne soit pas garanti). Il est cependant à prévoir que les grappes provenant de la partie mise à jour du cadre de recensement donnerait des listes s'écartant beaucoup de celles du recensement pour ce qui est aussi bien du nombre total de ménages que de leur identification spécifique.

49. Une dernière observation s'impose concernant l'utilisation d'un recensement ancien : il s'agit de la validité plutôt que de la variance de l'échantillon. Comme on l'a vu, les grappes sont habituellement sélectionnées en fonction d'une probabilité proportionnelle à la taille. Si la taille des grappes à croissance rapide n'est pas actualisée avant la sélection de l'échantillon, l'on risque une grave sous-représentation des zones qui ne comportaient qu'un petit nombre de ménages lors du recensement mais qui se sont beaucoup développées depuis lors. Les résultats de l'enquête seraient biaisés et évidemment trompeurs car les caractéristiques des personnes vivant dans de telles zones à forte croissance seront sans doute différentes de celles des personnes habitant dans des quartiers plus stables.

4.1.8.2. Cadres de recensement remontant à deux ans ou moins

50. La présente section s'applique aux pays qui, ayant réalisé des recensements assez récemment, c'est-à-dire au cours de l'année ou des deux années écoulées, n'auront pas besoin d'actualiser l'ensemble du cadre. En pareil cas, il serait sélectionné des grappes sur la *base de la taille* initialement définie lors du recensement car l'on peut la considérer comme assez exacte. La mise à jour en tant que telle n'interviendrait qu'à l'avant-dernière phase de la sélection, lorsque le personnel de terrain doit établir une liste à jour des ménages que comportent les grappes faisant partie de l'échantillon. Les ménages seraient sélectionnés sur la base des listes à jour et les pondérations seraient ajustées selon que de besoin conformément aux procédures examinées dans la section 3.7.2 dans le contexte de l'échantillonnage fondé sur une probabilité proportionnelle à la taille estimative.

51. S'il se peut qu'un petit nombre de grappes de l'univers se soient beaucoup développées depuis le recensement, la fréquence de ces cas ne sera généralement pas de nature à affecter beaucoup les opérations sur le terrain ou la précision de l'enquête. Si de telles grappes se trouvent faire partie de l'échantillon, elles peuvent être segmentées si besoin est. Cette segmentation, appelée « morcellement », est une procédure de terrain visant à alléger le travail représenté par l'établissement de listes. Cette procédure consiste à : *a*) diviser la grappe initiale en sections, habituellement des quadrants; *b*) à sélectionner une section au hasard aux fins de l'établissement d'une liste; et *c*) à sélectionner sur la base de ce segment les ménages qui devront être interrogés. Le morcellement n'améliore pas la fiabilité de l'échantillonnage car chaque segment doit se voir affecter un facteur de pondération supplémentaire égal au nombre de segments que comporte la grappe, c'est-à-dire quatre si celle-ci est divisée en quadrants. Néanmoins, le morcellement permet de maîtriser le coût des opérations sur le terrain. Un morcellement peut être nécessaire même si le recensement est récent, dans le cas également de zones d'énumération à forte croissance qui ont beaucoup changé depuis le recensement.

Il va de soi que, lorsque le recensement est extrêmement récent, il y aura généralement très peu de zones de ce type.

52. Il y a lieu de noter que des types d'inexactitudes précédemment mentionnées (ménages comptés deux fois ou ménages manquants, affectations à des zones d'énumération erronées) sont corrigées en partie lorsqu'il est établi, dans le contexte de l'opération de mise à jour, de nouvelles listes de ménages à l'avant-dernière phase. C'est là évidemment une autre raison qui milite solidement en faveur de l'établissement de listes à jour des ménages.

4.1.8.3. Cadre utilisé à une autre fin

53. Les responsables des enquêtes s'interrogent parfois sur le point de savoir si un cadre spécifiquement conçu pour un type d'enquête sur les ménages peut être utilisé pour une enquête d'un type différent. Par exemple, un cadre d'échantillonnage destiné à une enquête sur la population active peut-il être utilisé pour une enquête visant à mesurer l'état de santé d'une population, l'incidence des incapacités, la pauvreté ou la propriété de terres agricoles? Habituellement, cependant, ce n'est pas le cadre lui-même qui est problématique mais plutôt la façon dont il est stratifié. Le plus souvent, un cadre peut être utilisé pour des enquêtes différentes, à moins d'être incomplet, inexact ou obsolète dans le contexte des fins visées. L'on verra ci-après un exemple de cadres qui ne se prêtent pas à de multiples utilisations. Par exemple, si une enquête axée sur le coût de la vie vise uniquement les communautés urbaines (ce qui est fréquemment le cas dans la pratique), le cadre d'échantillonnage exclura les régions rurales. Manifestement, un tel cadre ne permettrait pas d'estimer la pauvreté dans les pays où celle-ci constitue essentiellement un phénomène rural.

54. Cependant, la plupart des enquêtes sur les ménages ont un caractère général du point de vue non seulement de leur contenu mais aussi de la conception des échantillons. Une enquête sur la population active vise généralement, par exemple, à rassembler des informations auxiliaires sur les caractéristiques démographiques, le niveau d'instruction et d'autres aspects. En pareil cas, la conception de l'échantillon la mieux appropriée est aussi une conception de caractère général, ce qui signifie qu'il y aura lieu d'utiliser un cadre habituel, c'est-à-dire un cadre qui couvre tous les ménages du pays. Il se peut que le cadre soit stratifié en fonction d'une variable propre à la mesure de la population active. Il se peut par exemple que les zones d'énumération soient classées selon le pourcentage de chômeurs lors du dernier recensement. L'on peut alors créer pour les zones d'énumération trois strates différentes : chômage modéré, moyen et élevé. Comme mentionné ci-dessus, c'est là une décision qui se rapporte à la stratification. Le cadre lui-même n'est pas affecté. La solution consisterait alors à « déstratifier » le cadre s'il doit être utilisé à d'autres fins, comme une enquête sanitaire.

55. L'une des principales tâches qui incombent aux statisticiens consiste à évaluer le cadre d'échantillonnage existant lorsqu'il doit être utilisé pour un autre type d'enquête. Cette évaluation doit notamment tendre à déterminer que le cadre existant correspond aux objectifs de mesure de l'enquête proposée, en particulier, comme souligné dans tout ce chapitre, en ce qui concerne la complétude, l'exactitude et l'actualité.

4.2. Cadres directeurs d'échantillonnage

56. Les échantillons-maîtres peuvent être d'un bon rapport coût-efficacité lorsque le pays mène un nombre suffisant d'enquêtes indépendantes ou de séries périodiques de la même enquête pour que leur utilisation se justifie. Il peut être évident que ces échantillons-maîtres doivent être conçus

comme il convient, mais il importe aussi au plus haut point de bien les tenir avec le temps. L'Organisation des Nations Unies (1986) a examiné beaucoup plus en détail la question des cadres directeurs d'échantillonnage et de leurs utilisations.

4.2.1. Définition et utilisation d'un échantillon-maître

57. Le cadre ou les cadres d'échantillonnage utilisés lors de la première phase de sélection doivent englober l'intégralité de la population cible. Lorsque ce cadre est utilisé pour des enquêtes multiples de multiples séries de la même enquête, il est appelé cadre directeur d'échantillonnage ou, plus simplement, échantillon-maître.

58. L'utilisation d'un cadre directeur est la stratégie privilégiée par tout pays qui mène un vaste programme continu d'enquêtes sur les ménages entre les différents recensements. Inversement, lorsqu'il n'existe pas de programmes continus d'enquêtes, des échantillons-maîtres ne sont généralement pas recommandés. Utiliser les mêmes cadres peut permettre de réaliser des économies d'échelle car, pour l'essentiel, c'est au stade de l'élaboration du cadre directeur qu'est encourue la majeure partie des coûts, plutôt que chaque fois que l'enquête est menée sur le terrain. D'un autre côté, ceux qui ne mènent qu'une enquête nationale occasionnelle entre deux recensements n'ont guère intérêt à utiliser un échantillon-maître.

59. Les caractéristiques d'un échantillon-maître tiennent au nombre, à la taille et au type d'unités sélectionnées lors de la première phase. D'une manière générale, un échantillon-maître se compose d'une sélection initiale d'unités primaires d'énumération (UPE) qui demeurent fixes pour chaque sous-échantillon. Il y a lieu de noter que les étapes suivantes sont généralement variables. Par exemple, à la dernière phase de la sélection, les ménages qu'il est décidé d'interroger sont habituellement différents lorsqu'il s'agit d'enquêtes indépendantes, mais peuvent être les mêmes lorsque l'on se trouve en présence d'enquêtes répétitives.

4.2.2. Caractéristiques idéales des unités primaires d'échantillonnage à retenir pour un cadre directeur

60. Les principes qui régissent l'établissement d'un cadre directeur sont très semblables à ceux qui s'appliquent aux cadres d'échantillonnage en général. Un cadre directeur doit être aussi complet, exact et à jour que possible et il ressemble beaucoup au cadre d'échantillonnage normalement élaboré pour le dernier recensement. Comme le cadre directeur peut être utilisé pendant l'intégralité de la période intercensitaire, cependant, il devra être mis à jour périodiquement, par exemple tous les deux ou trois ans, à la différence d'un cadre normal, qui n'est généralement mis à jour que de façon ponctuelle, seulement lorsqu'il est prévu de réaliser une enquête.

61. Comme l'on pouvait s'y attendre, les conditions qui doivent être réunies pour pouvoir élaborer un cadre directeur sont les mêmes que celles qui s'appliquent aux cadres d'échantillonnage en général. La définition des unités à utiliser comme UPE, par exemple, est subordonnée à la condition qu'elles soient déjà des zones définies sur une carte. Cette contrainte n'est cependant pas trop rigoureuse étant donné que les unités du cadre seront invariablement définies comme étant les unités administratives déjà construites pour le recensement. Une condition importante qui peut s'écarter de celles qui s'appliquent aux cadres d'échantillonnage ordinaires est cependant que la taille des UPE doit être suffisamment grande pour pouvoir réaliser de multiples enquêtes sans que les mêmes

déclarants doivent être interrogés à plusieurs reprises, mais même cette condition peut être assouplie dans certains cas.

Exemple

Un type spécifique de cadre directeur qui a été utilisé dans certaines situations est fondé sur une conception en deux phases. La première phase consiste à constituer un échantillon de nombreuses zones d'énumération (ou d'unités géographiques aussi petites). Il est sélectionné un sous-échantillon de l'échantillon-maître des zones d'énumération pour chaque enquête indépendante utilisant le cadre retenu. Chaque sous-échantillon donne lieu à l'établissement d'une liste ou est morcelé pour la préparation de l'enquête. Ainsi, l'échantillon-maître peut comporter 10 000 zones d'énumération, 1 000 sont retenues comme sous-échantillon en vue d'une enquête sur l'emploi. L'élaboration d'une liste des ménages est entreprise dans les 1 000 *zones d'énumération* à partir desquelles il est sélectionné pour l'enquête, lors d'une seconde phase, un échantillon de 15 ménages par zone d'énumération. L'année suivante, il est sélectionné à partir de l'échantillon-maître un autre sous-échantillon de 800 zones d'énumération qui sera utilisé pour une enquête sanitaire, et ainsi de suite. Ainsi, aucune *zone d'énumération* n'est utilisée plus d'une fois, de sorte que la taille de l'UPE est sans importance.

62. La taille de l'UPE est néanmoins importante lorsque tous les sous-échantillons générés à partir du cadre directeur doivent provenir de la même série d'UPE; dans l'exemple susmentionné, les zones d'énumération sont les *unités primaires d'échantillonnage,* et il est utilisé pour chaque sous-échantillon une sous-série différente de zones d'énumération. La sélection des UPE d'un cadre directeur ne pose pas de problème particulier car la méthode est la même qu'il s'agisse d'un échantillon-maître ou de tout autre échantillon. Généralement, il est appliqué la méthode de la probabilité proportionnelle à la taille, sauf dans de rares cas où les UPE sont de tailles plus ou moins égales; l'on peut alors utiliser un échantillon d'UPE sélectionnées sur la base d'une probabilité égale.

4.2.3. Utilisation d'échantillons-maîtres pour faciliter les enquêtes

63. L'on a vu dans la section 3.3.7 pourquoi un échantillon nombreux est nécessaire pour établir des échantillons-maîtres, afin de disposer d'un nombre suffisant de ménages pour réaliser de multiples enquêtes sur plusieurs années sans devoir interroger de façon répétée les mêmes déclarants. La taille prévisible des échantillons pour toutes les enquêtes qu'il est envisagé de réaliser avec le même cadre directeur constitue le principal paramètre à utiliser. Par exemple, s'il est prévu que 50 000 ménages seront interrogés lors des diverses enquêtes réalisées au moyen de l'échantillon-maître, les enquêteurs disposeront des informations essentielles dont ils ont besoin pour déterminer le nombre et la taille des *unités primaires d'échantillonnage*. De même, il peut être élaboré un plan d'exécution des enquêtes en se référant à l'échantillon-maître, comme le montre l'exemple suivant (voir aussi, à titre de comparaison, l'exemple donné dans la section 3.3.7).

Exemple

Comme dans l'exemple donné dans la section 3.3.7, l'échantillon-maître, dans le pays A, comprend 50 000 ménages. Il est prévu trois enquêtes pour lesquelles la taille des échantillons et des grappes doit être la suivante : 16 000 ménages et 6 ménages par grappe pour l'enquête sur les revenus et les dépenses; 12 000 ménages et 12 ménages par grappe pour l'enquête sur la population active; et 10 000 ménages et 20 ménages par grappe pour l'enquête sanitaire. Des grappes de tailles différentes sont sélectionnées pour tenir compte des effets différenciés (par type

d'enquête) du deff. En outre, il est maintenu en réserve 12 000 ménages pour réaliser d'autres enquêtes si besoin est. Il faudra, pour les trois enquêtes envisagées, 4 167 unités primaires d'échantillonnage (16 000/6 + 12 000/12 + 10,000/20). Comme le contenu des enquêtes pour lesquelles pourra être utilisé le sous-échantillon conservé en réserve est inconnu, il est décidé de prévoir une grappe de 12 ménages, ce qui ajoute 1 000 unités primaires d'échantillonnage de plus, et donne un total global de 5 167. L'équipe chargée de la conception de l'échantillon-maître décide par conséquent de construire un échantillon-maître de 5 200 unités primaires d'échantillonnage. La définition de l'unité primaire d'échantillonnage doit tenir compte du nombre de ménages à interroger. Dans cet exemple, chaque unité primaire d'échantillonnage doit être suffisamment grande pour que 50 ménages puissent être interrogés. Sur la base de cette information, les enquêteurs peuvent alors déterminer quelles sont les unités géographiques qui correspondent le mieux aux unités primaires d'échantillonnage. Si le pays A a des zones d'énumération comportant en moyenne une centaine de ménages, sans guère de variation de part et d'autre de cette moyenne, il pourra y avoir lieu d'utiliser des zones d'énumération comme unités primaire d'échantillonnage.

4.2.3.1. Avantages d'utilisations multiples d'un échantillon-maître

64. Utiliser un échantillon-maître présente de nets avantages. Premièrement et surtout, il constitue un outil de coordination pour les ministères fonctionnels et les autres organismes qui participent au programme national de statistiques. Cela vaut pour plusieurs aspects des enquêtes, indépendamment des considérations liées à l'échantillonnage en tant que tel : essentiellement maîtrise des coûts et élaboration de procédures normalisées au niveau des différents secteurs en ce qui concerne les définitions statistiques, le libellé des questions à poser lors des enquêtes et des procédures de codage des données.

65. L'un des principaux avantages d'un échantillon-maître est l'utilisation des mêmes unités primaires d'échantillonnage. Le personnel de terrain peut être organisé et maintenu en place aussi longtemps qu'est utilisé l'échantillon-maître. Par exemple, l'on peut recruter et former des enquêteurs de sorte qu'ils soient disponibles au début du programme couvert par l'échantillon-maître lorsque l'on sait où se trouvent les unités primaires d'échantillonnage qui seront utilisées pour toutes les enquêtes qui doivent être réalisées sur une période d'une dizaine d'années par exemple. Si besoin est, les enquêteurs peuvent être recrutés parmi les résidents de la localité où se trouvent les unités primaires d'échantillonnage que comprend l'échantillon-maître ou à proximité. Les documents nécessaires, comme les cartes des unités primaires d'échantillonnage et les listes de ménages peuvent être établis dès le début du programme d'enquêtes devant utiliser l'échantillon-maître, ce qui permettra de gagner du temps ainsi que d'amortir une proportion significative des coûts sur l'ensemble des enquêtes prévues. En outre, le fait que l'échantillon sera utilisé plusieurs fois offre une possibilité de mieux contrôler les erreurs autres que les erreurs d'échantillonnage et même les non-réponses. En effet, rendre visite plusieurs fois aux mêmes déclarants peut permettre de prendre note de leurs attitudes et d'identifier les problèmes qui se posent dans différents domaines, ce qui permettra de mettre au point des mesures correctives pour les enquêtes suivantes. Cependant, il importe d'insister à nouveau sur le fait que de tels avantages apparaissent seulement lorsque l'échantillon-maître doit être très largement utilisé.

66. De manière générale, les autres avantages des échantillons-maîtres, que l'on utilise les mêmes unités primaires d'échantillonnage dans le cadre d'une conception en trois phases ou des unités

primaires d'échantillonnage différentes dans une conception à deux phases, sont notamment la possibilité : *a*) d'intégrer, aux fins de l'analyse, des données provenant de deux ou plusieurs utilisations de l'échantillon-maître avec un contenu différent; et *b*) d'intervenir rapidement lorsque surgit la nécessité de rassembler des données imprévues.

4.2.3.2. Limitations d'utilisations multiples d'un cadre directeur d'échantillonnage

67. Il peut y avoir certaines limitations à l'utilisation d'un échantillon-maître, par exemple le risque d'épuiser la réserve d'unités primaires d'échantillonnage, c'est-à-dire de ménages, s'il est utilisé trop fréquemment. L'on peut cependant remédier à cette difficulté au moyen d'une planification appropriée. Bien que l'on ne puisse pas prévoir toutes les utilisations qui pourront être faites d'un échantillon-maître pendant tout son cycle de vie utile, l'on peut, si l'échantillon-maître est suffisamment nombreux, mettre en réserve des sous-sous-échantillons en vue d'autres utilisations possibles.

68. Une autre limitation tient aux distorsions toujours plus grandes qui apparaissent lorsque l'échantillon-maître n'est pas mis à jour comme il convient à mesure qu'il vieillit. Enfin, un échantillon-maître se prête mal à la collecte de données dans le contexte d'enquêtes « spéciales » comme celles qui portent sur des provinces spécifiques ou des sous-populations rares.

4.2.4. Allocation entre les différents domaines (régions administratives, etc.)

69. Il est de plus en plus fréquemment demandé aux offices nationaux de statistiques de tabuler et d'analyser les données provenant des enquêtes sur les ménages pour d'importantes circonscriptions administratives infranationales comme régions, provinces et grandes villes. Certains pays, comme le Viet Nam, sont même censés publier systématiquement des données au niveau des districts. Ces exigences et ces attentes répondent à des besoins légitimes des pouvoirs publics, essentiellement parce que les programmes socioéconomiques sont axés sur des régions localisées plutôt que sur le pays dans son ensemble.

70. Dans le jargon de l'échantillonnage, comme on l'a vu, ces circonscriptions constituent des domaines. Comme la taille de l'échantillon nécessaire pour obtenir des résultats fiables est énorme, ce type d'échantillonnage suppose des coûts qui dépassent fréquemment les crédits que les gouvernements peuvent dégager pour des enquêtes. La nécessité pour recueillir des données au niveau des domaines affecte également l'élaboration du cadre directeur d'échantillonnage.

71. Le nombre et le type de domaines à établir ont été examinés dans la section 3.3.4 concernant la taille des échantillons, et l'on ne reviendra pas ici sur cette question. Une fois que ces décisions ont été prises, l'on peut construire le cadre directeur d'échantillonnage. Par exemple, il se peut qu'un pays décide que les enquêtes prévues par le programme ne devront rassembler de données que pour deux domaines, les régions urbaines et les régions rurales. Un autre pays comptant 12 provinces souhaitant obtenir des estimations pour chacune d'elles pourra décider que les ressources disponibles permettent de prendre des échantillons suffisamment nombreux pour que les provinces puissent être traitées comme des domaines. Un troisième pays comptant 50 provinces pourra, quant à lui, décider qu'il est trop coûteux de chercher à obtenir des estimations pour chacune d'elles et choisir de définir comme domaine les huit grandes régions géographiques qui regroupent les 50 provinces. Un quatrième pays pourra décider de ne pas établir de domaines en tant que tels mais plutôt de tabuler l'échantillon national, alloué proportionnellement par région, par province, par circonscription urbaine, par circonscription rurale et pour quelques grandes villes sélectionnées, l'intention étant

de publier les données concernant les sous-zones pour lesquelles la taille de l'échantillon est jugée suffisante pour donner des résultats raisonnablement fiables.

72. Dans le cas des exemples ci-dessus, l'allocation des domaines n'intervient pas étant donné que l'échantillon est réparti proportionnellement entre les différentes zones infranationales visées. Dans le cas des trois premiers exemples, il faut s'attacher particulièrement à allouer comme il convient les UPE de l'échantillon-maître. Comme l'estimation des domaines implique une fiabilité égale pour chacune des sous-populations ou zones définies comme étant un domaine, il sera sélectionné dans chacun d'eux le même nombre d'UPE, règle qui s'applique sans égard à la question de savoir si la conception de l'échantillon est fondée ou non sur un échantillon-maître.

4.2.5. Maintenance et mise à jour des échantillons-maîtres

73. Eu égard à l'impact qu'il a sur la couverture de la population, il importe au plus haut point de prévoir, lors de son élaboration, une maintenance appropriée du cadre directeur d'échantillonnage. L'échantillon-maître d'un pays donné est habituellement utilisé pendant les dix années intercensitaires, période pendant laquelle il y a généralement d'importants changements démographiques. Il faut par conséquent mettre à jour périodiquement le cadre afin de refléter ces changements et faire en sorte qu'ils demeurent « représentatifs ».

74. Deux types de mise à jour sont importants. Le premier, le plus simple, consiste à préparer de nouvelles listes de ménages pour les grappes sélectionnées à l'avant-dernière phase, procédure qui est généralement recommandée dans l'ensemble de ce guide, qu'il s'agisse des échantillons-maîtres ou des échantillons à usage unique. Les grappes sont ainsi automatiquement mises à jour pour refléter les migrations, les naissances et les décès. Ce type de mise à jour, qui porte uniquement sur les grappes, aide à minimiser les erreurs de couverture (autres que d'échantillonnage), mais la variance de l'échantillonnage augmente avec le temps si l'ensemble du cadre n'est pas mis à jour.

75. Il faut également mettre à jour périodiquement le cadre d'échantillonnage dans son ensemble de manière à refléter comme il convient, à grande échelle, la croissance post-censitaire. Comme on l'a vu, cette croissance se produit généralement lorsque sont construits de grands immeubles résidentiels et lorsque se développent des populations urbaines spontanées. Les zones d'énumération caractérisées par une telle croissance étaient invariablement beaucoup plus petites lorsque le cadre directeur a été établi. De ce fait, les tailles sélectionnées se traduisent par une grave sous-estimation, ce qui réduit leurs chances de sélection dans le cas d'une conception fondée sur la probabilité proportionnelle à la taille. L'effet sur la variance de l'échantillonnage peut être considérable lorsque de telles zones d'énumération se trouvent effectivement sélectionnées étant donné que leur taille actuelle peut être plusieurs fois supérieure à ce qu'elle était initialement.

76. Les problèmes que soulèvent les zones à forte croissance et leur impact sur les échantillons-maîtres peuvent être considérablement réduits en mettant à jour le cadre périodiquement, par exemple, tous les deux ou trois ans.

4.2.6. Remplacement par roulement des UPE dans les échantillons-maîtres

77. Le lecteur est invité à se référer à la section 3.9.2 du chapitre 3, intitulée « Échantillonnage visant à estimer un changement ou une tendance » pour une discussion détaillée du chevauchement des échantillons dans le cas d'enquêtes répétitives ou continues visant à mesurer un changement ou

une tendance. Un chevauchement implique l'existence d'un système d'échantillonnage prévoyant des ménages de remplacement lorsque les enquêtes sont répétées. Il importe de souligner à nouveau que l'utilisation d'échantillons qui se chevauchent est la méthode privilégiée pour estimer le changement intervenu, par exemple, d'une année sur l'autre. Une méthode pouvant être envisagée consiste à remplacer certains éléments de l'échantillon par roulement, ce qui se traduit par un chevauchement partiel d'une enquête ou d'une occasion à l'autre.

78. L'on a fait observer dans la section précédente qu'aussi bien la fiabilité de l'échantillonnage (élément positif) que les erreurs autres que d'échantillonnage (élément non positif) sont les plus grandes lorsque ce sont les mêmes ménages qui sont utilisés pour chaque série de l'enquête. De ce fait, il faut souvent rechercher un compromis par le biais d'un chevauchement partiel de l'échantillon d'une série à l'autre, surtout lorsqu'une enquête est répétée trois fois ou plus (voir la section 3.9.2 pour la raison d'être d'un chevauchement partiel).

79. Pour introduire un chevauchement partiel, une méthode consiste à remplacer les UPE par roulement (par opposition à un remplacement des ménages faisant partie de l'échantillon d'UPE). Lorsqu'un échantillon-maître d'UPE est utilisé non seulement pour plusieurs séries d'une même enquête mais pour de multiples enquêtes, il est tout aussi important d'étudier soigneusement la nécessité de prévoir un roulement.

80. Pour qu'un plan de roulement puisse être appliqué dans la pratique et donne des résultats acceptables, le degré de chevauchement d'une période à l'autre doit être identique et constant avec le temps. Par exemple, si le chevauchement entre les années 1 et 2 est de K %, il devra également être de K % entre les années 2 et 3, entre les années 3 et 4, et ainsi de suite. En conséquence, il faut intégrer cet aspect pour préparer le plan de roulement de toutes les UPE.

4.2.6.1. Exemples nationaux d'échantillons-maîtres

81. L'on trouvera ci-après une description des échantillons-maîtres utilisés dans quatre pays en développement : le Cambodge, les Émirats arabes unis, le Viet Nam et le Mozambique. Chaque exemple illustre certaines des caractéristiques des échantillons-maîtres et certains des principes applicables, comme l'échantillonnage en une ou deux phases des unités devant constituer l'échantillon-maître et son application flexible dans le cas d'enquêtes spécifiques. La présente section illustre également d'autres aspects de la conception de l'échantillon dont il est question dans le guide, et notamment la stratification implicite, le choix optimal de la taille des grappes pour réduire les effets du *deff* et l'allocation de l'échantillon entre différents domaines.

4.2.6.2. Cambodge, 1998-1999

82. Le plan directeur du Cambodge illustre l'utilisation d'une conception à deux phases. Un échantillon d'UPE de grande taille a donné une liste des segments à sélectionner lors d'une seconde phase en vue d'enquêtes spécifiques.

83. L'Institut national cambodgien de la statistique a élaboré en 1999 un échantillon-maître devant être utilisé pour le Programme intercensitaire d'enquêtes sur les ménages, qui comprend une enquête socioéconomique périodique et, le cas échéant, des enquêtes sur la santé, la population active, les revenus et les dépenses et la démographie ainsi que des enquêtes ad hoc. Le recensement de la population de 1997 a constitué le cadre qui a servi à la conception de l'échantillon-maître, laquelle a été réalisée en deux phases. La première phase a consisté à sélectionner les UPE sur la base d'une

probabilité proportionnelle à la taille, les UPE étant définies comme les villages d'une certaine taille, selon le dénombrement des ménages effectué lors du recensement. La sélection des UPE s'est faite par ordinateur. La seconde phase a consisté à créer des segments à l'intérieur des UPE sélectionnées, opération qui a été réalisée manuellement.

84. Il a été décidé d'utiliser les villages comme UPE car ils étaient de taille suffisante, comportant en moyenne assez de ménages (245 en milieu urbain et 155 en milieu rural) pour pouvoir procéder à plusieurs enquêtes pendant la période intercensitaire. Ainsi, il ne serait pas nécessaire d'interroger plusieurs fois les mêmes déclarants. La possibilité d'utiliser les zones d'énumération du recensement a été envisagée mais écartée car, en moyenne, leur taille n'était que la moitié de celle des villages. Certains groupes particuliers, comme des populations itinérantes ou vivant en établissement, n'ont pas été inclus dans l'échantillon-maître, pas plus que les casernes.

85. Il a été sélectionné au total 600 UPE pour l'échantillon-maître car l'on a estimé que ce chiffre serait suffisamment réparti sur l'ensemble du pays pour représenter comme il convient toutes les provinces. Une stratification implicite a été utilisée pour sélectionner l'échantillon : le fichier de villages a été trié dans l'ordre géographique, milieu urbain, milieu rural, province, district et commune. Ainsi, l'échantillon-maître a été alloué proportionnellement, automatiquement, par milieu urbain et rural et par province.

86. Une caractéristique intéressante de l'échantillon-maître élaboré au Cambodge a été la seconde phase de l'opération d'échantillonnage. Comme on l'a dit, il a été créé à cette fin des segments à l'intérieur des UPE sélectionnées. Il y a lieu de faire observer que le concept à la base de la seconde phase de la construction d'un échantillon-maître ne doit pas être confondu avec celui qui sous-tend la seconde phase de la sélection de l'échantillon, qui a trait à la sélection de ménages en vue d'enquêtes spécifiques. Il a été constitué à l'intérieur de chaque UPE sélectionnée des segments d'une dizaine de ménages en moyenne, ce qui n'a exigé aucun travail sur le terrain, sauf dans des cas exceptionnels, étant donné que l'on a pu utiliser les listes et les cartes établies pour le recensement de 1997. Le nombre de segments créés à l'intérieur de chaque UPE a été calculé comme étant le nombre de ménages dénombrés lors du recensement divisé par 10 et arrondi au nombre entier le plus proche. Par exemple, un village contenant 187 ménages selon le recensement de 1997 a été divisé en 19 segments.

87. Les segments ainsi créés ont constitué la source des échantillons ou sous-échantillons à utiliser pour différentes enquêtes. Un ou plusieurs segments sont choisis parmi l'intégralité ou une partie des UPE pour chaque enquête ou chaque série d'enquêtes utilisant l'échantillon-maître. Un avantage important est que la création de l'échantillon-maître comme indiqué donne aux différentes enquêtes qui l'utilisent la possibilité d'être autopondérées, selon la conception de l'échantillon.

88. L'un des principaux avantages de l'échantillon-maître est qu'il permet une grande souplesse quant aux sous-échantillons à sélectionner pour les enquêtes spécifiques. La sélection des grappes (c'est-à-dire des segments) pour chaque enquête peut donner une série différente si on le souhaite. L'UPE type contient environ 18 à 30 segments, de sorte que le nombre de segments que comporte chaque UPE est suffisant pour pouvoir réaliser toutes les enquêtes. De plus, une enquête socioéconomique peut être répétée chaque année avec une série différente de segments. À défaut, l'on peut également opter pour un chevauchement de l'échantillon en conservant certains des segments (par exemple 25 %) d'une année sur l'autre, par roulement, 75 % des segments étant remplacés chaque année.

89. Une des limitations de l'échantillon-maître tient à l'utilisation de grappes compactes (tous les ménages du segment pris comme échantillon sont adjacents), ce qui accroît quelque peu l'effet de conception par rapport à ce qu'il est dans le cas de segments non compacts, c'est-à-dire d'un échantillon systématique de ménages d'une grappe de plus grande taille étant donné que, apparemment, l'effet de conception pourrait être réduit jusqu'à un certain point en limitant la taille des grappes, c'est-à-dire des segments, à dix ménages seulement.

90. L'on prévoyait de mettre à jour l'échantillon tous les deux ou trois ans. Il était admis qu'il serait préférable de mettre à jour l'intégralité de l'échantillon-maître, mais il a été décidé de n'actualiser les UPE que pour l'échantillon devant servir à l'enquête spécifique alors envisagée. Pour mettre à jour l'échantillon, il a été organisé des visites sur le terrain pour établir une nouvelle liste des ménages se trouvant dans les segments affectés. Cette nouvelle liste a été établie pour les mêmes zones géographiques où se trouvait la série initiale de ménages, et c'est l'une des raisons pour lesquelles les limites géographiques des segments présentent tant d'importance.

91. Il est intéressant de noter que les notables des villages ont été appelés à participer aux opérations de mise à jour. L'on sait en effet qu'ils tiennent des registres soigneux de tous les ménages de leurs villages, et ces registres sont systématiquement tenus à jour. Le plus souvent, les listes ainsi tenues sont considérées comme fort exactes. En outre, les notables des villages ont également été des sources d'informations précieuses pour identifier et délimiter les zones géographiques correspondant aux différents segments.

4.2.6.3. Émirats arabes unis, 1999

92. L'échantillon-maître établi aux Émirats arabes unis présente, par sa conception, deux caractéristiques notables. Premièrement, l'échantillon-maître a été conçu avec une stratification spéciale pour tenir compte de l'existence, aux Émirats arabes unis, de deux populations diverses, à savoir les nationaux et les étrangers. Deuxièmement, la conception de l'échantillon illustre comment la conception du segment type (voir la section 3.8.2) peut être exploitée lorsque l'on se trouve en présence d'un recensement ancien et de zones d'énumération de tailles différentes.

93. L'échantillon-maître est décrit par le Ministère de la planification comme étant un super échantillon de 500 UPE fondé sur le recensement de la population de 1995, pris comme cadre directeur. Cet échantillon est censé être utilisé pour différentes enquêtes sur les ménages jusqu'à la réalisation du prochain recensement. Les UPE sont définies comme étant les zones d'énumération ou des parties des zones d'énumération du recensement, de sorte qu'en moyenne chaque UPE contient environ 60 ménages, aussi bien nationaux qu'étrangers.

94. Il a été construit deux strates avant de sélectionner les UPE. La première comprenait les zones d'énumération dans lesquelles, au moment du recensement, un tiers ou plus des ménages étaient des nationaux. La seconde strate comprenait toutes les autres zones d'énumération. Il a été rangé 1 686 zones d'énumération dans la première strate et 2 986 dans la seconde. Il a alors été sélectionné, sur la base d'une probabilité proportionnelle à la taille, un échantillon de 250 UPE, ce qui a donné au total 500 UPE pour l'ensemble du pays. L'idée était que cet échantillon-maître donnerait un nombre à peu près égal de ménages nationaux et de ménages étrangers. Premièrement, les UPE de grande taille (comportant 90 ménages ou plus) ont été morcelées et un segment a été choisi au hasard. Une nouvelle liste actualisée des ménages a été établie pour les 500 UPE pour mettre à jour le cadre d'échantillonnage. L'on a ainsi obtenu environ 30 000 ménages pouvant être utilisés selon des

combinaisons diverses pour des enquêtes spécifiques. Dans un souci de souplesse, les ménages que contenait chaque UPE de l'échantillon ont été subdivisés en 12 sous-séries de 5 ménages chacune.

95. Un aspect notable de l'échantillon-maître est qu'il n'est pas autopondéré car les deux strates ne sont pas de même taille. L'échantillon-maître a été utilisé pour la première fois lors de l'Enquête nationale de 1999 sur le diabète, menée sous l'égide du Ministère de la santé. Il était également prévu de réaliser une enquête sur la population active et une enquête sur les revenus et les dépenses (ou enquête sur le budget des ménages).

96. L'on trouvera ci-après plusieurs autres détails sur certaines circonstances spéciales qui ont déterminé la conception de l'échantillon-maître. Deux considération prédominantes dont il a été soigneusement tenu compte ont été les populations cibles et le cadre d'échantillonnage.

97. Comme on l'a dit, il y avait dans le pays deux importantes populations cibles, les nationaux et les étrangers. Les premiers constituaient environ 43 % de la population, selon le recensement de la population de 1995, mais un quart seulement des ménages du pays. En effet, les ménages étrangers comptaient chacun beaucoup moins de membres. Il en résultait, pour la conception de l'échantillon, que s'il était sélectionné un échantillon proportionnel des ménages du pays, près des trois quarts des ménages interrogés seraient étrangers. Une autre conséquence était que la fiabilité des estimations pour les ménages étrangers serait environ trois fois plus grande que pour les ménages nationaux. Comme les résultats de l'enquête devaient servir à la formulation des politiques et des programmes nationaux, une telle disparité dans la fiabilité des estimations était à éviter. La solution, du point de vue de la conception de l'échantillon, a consisté à traiter les deux populations cibles disparates et inégales comme des entités distinctes en procédant à une stratification appropriée, comme décrit ci-dessus.

98. Il a également été utilisé un autre niveau de stratification, à savoir une stratification géographique, afin de répartir comme il convient l'échantillon par émirat et par régions urbaine et rurale. Le fichier des zones d'énumération a été trié dans l'ordre suivant avant la sélection d'échantillons : premièrement, la strate nationale par région urbaine et, à l'intérieur des régions urbaines, par émirat, puis, par émirat, par codes de zone d'énumération, dans l'ordre ascendant, selon le pourcentage de nationaux, ensuite par région rurale, par émirat et par code de zone d'énumération, et enfin la strate étrangère, dans le même ordre.

99. Il était admis que l'une des principales caractéristiques du cadre directeur d'échantillonnage devait être une série claire de cartes définissant les zones devant être couvertes par l'échantillon, c'est-à-dire les unités primaires d'échantillonnage. Les zones géographiques devaient être suffisamment petites pour qu'une liste puisse facilement être établie, mais en même temps assez grandes de manière à pouvoir être clairement définies par des frontières naturelles (afin de pouvoir être localisées plus facilement). Les zones d'énumération apparaissaient comme les seules zones répondant à ce double critère. Regrettablement, il n'avait pas été utilisé de cartes lors de recensements de la population, de sorte que les zones d'énumération existantes n'étaient pas clairement définies par des limites connues. De ce fait, il a fallu rassembler des informations précises sur les limites des UPE, c'est-à-dire des zones d'énumération.

100. Lors de la préparation de l'échantillon-maître d'UPE, l'on a utilisé la « conception du segment type » décrite au chapitre 3. Il s'agit d'une méthodologie qui a été utilisée avec succès dans de nombreux pays dans le cadre du Programme d'enquêtes démographiques et sanitaires et de l'Enquête panarabe sur la santé maternelle et infantile (PAPCHILD).

101. Il a été décidé d'utiliser la méthode du segment type car, aux Émirats arabes unis, les zones d'énumération du recensement sont de tailles très variables. Des segments types d'environ 60 ménages ont été créés. Le nombre de segments types a été calculé pour chaque UPE comme étant le nombre total de ménages (c'est-à-dire de ménages nationaux et étrangers) divisé par 60, le résultat étant arrondi au nombre entier le plus proche.

102. Lorsque le nombre de segments, c'est-à-dire la taille, était égal ou supérieur à deux, l'UPE a été subdivisée en segments. Il a fallu pour cela visiter la zone d'énumération (UPE) et établir une carte au moyen d'un dénombrement rapide et d'une localisation des logements (et non des ménages). Après morcellement, il a été choisi un segment au hasard dans chaque UPE. Ce segment est devenu la zone géographique effectivement utilisée pour l'échantillon-maître. Il a été organisé une autre visite sur le terrain pour établir une liste à jour des ménages de chaque segment pris comme échantillon. Cette dernière opération, nécessaire pour mettre à jour le cadre directeur, qui remontait à trois ans, a été considérée comme un aspect d'importance capitale de l'opération d'échantillonnage.

103. La dernière opération a consisté à subdiviser les ménages figurant sur la nouvelle liste en 12 sous-séries ou panels systématiques. Un ou plusieurs des panels devaient être utilisés pour des enquêtes spécifiques. Comme la taille moyenne du segment était d'une soixantaine de ménages, chaque panel contenait en moyenne 5 ménages.

104. S'il a été décidé d'avoir 12 panels, c'était essentiellement en raison de la souplesse qu'offrait ce chiffre s'agissant de combiner les panels à utiliser lors des enquêtes. Le choix effectivement fait pour une enquête donnée devait dépendre de différents facteurs, dont les objectifs de l'enquête, la taille souhaitée des grappes et la taille de l'échantillon global requis. Par exemple, deux cinquièmes des UPE devaient être utilisées pour l'Enquête nationale sur le diabète. Il a donc été pris 4 des 12 panels de ménages de ces UPE. Cette combinaison a donné un plan directeur global de 200 UPE avec des grappes de 20 ménages (c'est-à-dire quatre fois 5 ménages) et un échantillon global de quelque 4 000 ménages.

4.2.6.4. Viet Nam, 2001

105. L'échantillon-maître utilisé au Viet Nam présente deux caractères distinctifs. Il démontre l'utilisation de deux phases pour la sélection de l'échantillon-maître et d'une troisième lors de son application à des enquêtes ponctuelles. Deuxièmement, il démontre comment un échantillon-maître peut être alloué à des domaines géographiques.

106. L'échantillon-maître, qui est fondé sur le cadre directeur constitué par le recensement de 1999, repose sur une conception en deux phases. Les UPE ont été définies comme étant les communes dans les régions rurales et les quartiers dans les régions urbaines. Elles ont été ainsi définies car il a été décidé qu'il faudrait que chaque UPE comporte au minimum 300 ménages pour pouvoir être utilisée pour l'échantillon-maître. Il a également été envisagé de prendre comme UPE les zones d'énumération, mais celles-ci étaient trop petites et auraient dû être combinées avec des zones d'énumération adjacentes pour constituer de bonnes UPE. Ce travail a été jugé trop long et trop fastidieux. D'un autre côté, le nombre de communes et de quartiers qui ont dû être combinés en raison de leur petite taille n'a été que de 529 sur plus de 10 000.

107. Il a été sélectionné au total 3 000 UPE en fonction d'une probabilité proportionnelle à la taille. Chaque UPE contenait en moyenne 25 zones d'énumération en milieu urbain et 14 en milieu rural. Pour la seconde phase de la sélection, il a été retenu trois zones d'énumération dans chaque

UPE retenue pour l'échantillon sur la base d'une probabilité proportionnelle à la taille. Les unités secondaires, les zones d'énumération, contenaient en moyenne une centaine de ménages selon le recensement de 1999 : 105 dans les régions urbaines et 99 dans les régions rurales.

108. L'un des objectifs de l'échantillon-maître était de pouvoir rassembler des données passablement fiables pour chacune des huit régions géographiques du Viet Nam. La sélection de l'échantillon a été effectuée indépendamment dans chaque province. Ainsi, les provinces ont fait fonction de strates pour l'échantillon-maître. L'on voulait sursélectionner l'échantillon dans certaines provinces contenant des populations très peu nombreuses. En conséquence, l'allocation de l'échantillon entre les provinces a été faite selon la méthode de la probabilité proportionnelle à la racine carrée de la taille de la province. Il a été procédé à une allocation proportionnelle entre les régions urbaines et les régions rurales.

109. Indépendamment de la stratification au niveau provincial, l'on a eu recours à une stratification géographique implicite à l'intérieur des provinces. Pour utiliser l'échantillon-maître au cours des enquêtes spécifiques, l'on devait employer des sous-séries de zones d'énumération, par exemple un tiers d'entre elles pour l'Enquête générale sur les ménages. Lorsqu'il est prévu une enquête, il est administré une troisième phase de sélection, un nombre fixe de ménages étant choisi dans chaque zone d'énumération. Ce nombre peut varier selon l'enquête et selon qu'il s'agit d'une région urbaine ou rurale. Par exemple, l'on pourrait retenir un ménage par zone d'énumération pour des zones d'énumération rurales et 10 pour des zones d'énumération urbaines.

4.2.6.5. Mozambique, 1998-1999

110. L'échantillon-maître utilisé au Mozambique illustre l'utilisation d'une sélection en une seule phase d'UPE pour toutes les enquêtes nationales prévues par le Programme intercensitaire d'enquêtes sur les ménages. Il illustre également comment un échantillon-maître souple peut être adapté à la lumière des objectifs de mesure visés par telle ou telle enquête.

111. Au Mozambique, les UPE ont été définies et sélectionnées de façon simple, comme décrit dans divers chapitres du présent guide. Les UPE ont été construites à partir du cadre directeur constitué par le recensement de la population de 1997. Elles étaient composées de groupements géographiques comprenant en général de trois à sept zones d'énumération du recensement contenant en moyenne une centaine de ménages. Les UPE de l'échantillon-maître ont été sélectionnées sur la base d'une probabilité proportionnelle à la taille.

112. Il a été sélectionné au total 1 511 UPE pour constituer le cadre d'échantillonnage à utiliser dans le contexte du système intégré d'enquêtes sur les ménages du Mozambique. Les UPE ont été divisées en panels constituant chacun une sous-série systématique et par conséquent, en soi, un échantillon probabiliste. Il y a 10 panels comportant chacun 151 UPE. Dans le contexte du plan quinquennal (2000-2004), le Questionnaire sur les indicateurs du bien-être conçu par la Banque mondiale a été un instrument utilisé pour la première enquête menée avec l'échantillon-maître ainsi établi.

113. Le plan directeur de l'enquête a été conçu en ayant en vue deux objectifs : le premier était d'obtenir les indicateurs pertinents pour analyser le bien-être des personnes et des ménages au Mozambique. Le second objectif était d'obtenir des estimations fiables de ces indicateurs au plan national, séparément pour les régions urbaines et les régions rurales et pour chacune des 11 provinces du pays. La méthode d'échantillonnage utilisée pour l'enquête a été fondée sur le choix d'environ

14 500 ménages d'une conception stratifiée en grappes. En conséquence, la première phase de la sélection a évidemment porté sur les UPE de l'échantillon-maître.

114. La seconde phase de la sélection a porté sur un sous-échantillon des UPE de l'échantillon-maître. Il a été sélectionné au total 675 UPE sur les 1 511 de l'échantillon-maître. Les UPE ont été sélectionnées systématiquement sur la base d'une probabilité égale et, en outre, elles ont été allouées également entre les 11 provinces du Mozambique. Lors de la troisième phase, il a été sélectionné un échantillon d'une zone d'énumération dans chacune des UPE. Ainsi, l'échantillon retenu comportait 675 grappes : 475 rurales et 200 urbaines. Les zones d'énumération ont été sélectionnées sur la base d'une probabilité égale car leur taille était à peu près égale, comme mentionné ci-dessus, c'est-à-dire d'une centaine de ménages en moyenne. La dernière phase de la sélection a été réalisée à la suite de visites sur le terrain de l'Instituto Nacional de Estatistica (INE) afin d'établir une nouvelle liste de ménages pour mettre à jour le cadre directeur de 1997. À partir des listes ainsi compilées, il a été sélectionné pour l'enquête un échantillon systématique de 20 ménages dans les régions rurales et de 25 ménages dans les régions urbaines. La sélection de l'échantillon utilisé pour l'enquête s'est par conséquent faite en quatre phases, bien que l'échantillon-maître utilisé ait été conçu en une seule phase.

115. Deux caractéristiques de l'enquête sur les indicateurs du bien-être illustrent la souplesse avec laquelle l'échantillon-maître peut être adapté aux exigences spécifiques de l'enquête dont il s'agit.

116. Premièrement, pour utiliser l'échantillon-maître en vue de l'enquête sur les indicateurs du bien-être, l'Instituto Nacional de Estatistica souhaitait utiliser les panels déjà désignés, comme indiqué ci-dessus. Comme on voulait avoir pour l'enquête environ 600 UPE, l'on espérait que quatre des panels pourraient être utilisés. Cette idée a cependant été écartée lorsque l'on s'est rendu compte que le nombre de panels qui seraient nécessaires pour l'enquête varierait d'une province à l'autre étant donné que l'objectif de mesure exigeait un échantillon de taille plus ou moins égale dans chaque province. L'on a donc sélectionné les 675 UPE nécessaires, de manière systématique, en se reportant à l'ensemble de l'échantillon-maître, sans avoir recours à des panels. Si l'on s'est ainsi écarté de l'intention initiale de l'échantillon-maître, qui était de fournir un échantillon proportionnel par province, c'est parce que l'on souhaitait que les informations rassemblées au niveau des provinces soient de fiabilité égale. Selon le plan directeur tel qu'initialement conçu, les estimations au niveau national devaient primer.

117. Un deuxième aspect important de la conception de l'échantillon tenait à la taille des grappes. Il a été convenu que les grappes devraient être de tailles différentes pour les ménages urbains et les ménages ruraux, l'idée étant que l'effet dû à la conception de l'échantillon, c'est-à-dire le *deff,* était plus marqué dans les régions rurales, où l'agriculture de subsistance constituait la principale activité. Autrement dit, il est probable que les ménages des zones rurales présenteraient des caractéristiques très semblables. L'échantillon-maître offrait la possibilité de sélectionner un nombre différent de ménages (25 et 20, respectivement, dans les régions urbaines et rurales) lors de la dernière phase.

4.3. Résumé des lignes directrices à suivre

118. La présente section résume les principales lignes directrices pouvant être tirées de ce chapitre. Comme dans le cas du chapitre 3, il s'agit d'indications plutôt que de recommandations fermes. Ainsi, il convient :

- D'utiliser des cadres d'échantillonnage qui soient aussi complets, exacts et à jour que possible.
- De veiller à ce que le cadre d'échantillonnage couvre la population cible visée.
- D'utiliser le recensement le plus récent comme cadre directeur pour les enquêtes sur les ménages, si possible.
- De définir les UPE du cadre sous forme d'unités géographiques, comme les zones d'énumération du recensement, ayant des limites bien définies et comportant une population dont les effectifs sont connus.
- D'utiliser la liste des ménages établie lors du recensement comme cadre, lors de la dernière phase, seulement si elle est très récente, c'est-à-dire si elle n'a pas plus d'un an.
- D'utiliser avec prudence les doubles cadres ou les cadres multiples en veillant à mettre en place des procédures pour éviter les doubles emplois.
- De mettre à jour le cadre de recensement s'il remonte à plus de deux ans pour l'ensemble du pays ou dans des zones spécifiques à forte croissance et, à cette fin :
  - D'utiliser un quadrillage rapide pour mettre à jour l'ancien cadre.
  - De mettre à jour les grappes en établissant une nouvelle liste des ménages.
- De ne mettre à jour que les grappes si le cadre du recensement n'a pas plus de deux ans, et notamment :
  - D'établir une nouvelle liste de ménages.
- De n'utiliser l'échantillon-maître ou un cadre directeur que lorsqu'il est prévu ou a été entrepris un programme continu d'enquêtes de grande envergure.
- De définir les UPE de l'échantillon-maître de sorte qu'elles soient assez grandes ou assez nombreuses pour pouvoir être utilisées pour plusieurs enquêtes, ou des séries répétées de la même enquête, pendant la période intercensitaire.
- De mettre à jour les cadres directeurs en se référant aux mêmes lignes directrices que celles qui sont suggérées ci-dessus pour les cadres destinés à une seule enquête.
- D'utiliser la méthode de roulement des échantillons, par ménages ou par UPE, pour les enquêtes répétées utilisant des échantillons-maîtres.

Références et autres lectures

Banque mondiale (1999). *Core Welfare Indicators Questionnaire (CWIQ) Handbook.* Washington, Banque mondiale, chapitre 4.

Cochran, W. G. (1977). *Sampling Techniques,* troisième édition. New York, Wiley.

Fonds des Nations Unies pour l'enfance (2000). *End-Decade Multiple Indicator Survey Manual.* New York, UNICEF, chap. 4.

Hansen, M. H., W. N. Hurwitz et W. G. Madow (1953). *Sample Survey Methods and Theory.* New York, Wiley.

International Statistical Institute (1975). *Manual on Sample Design*. World Fertility Survey Basic Documentation, Voorburg (Pays-Bas).

Kalton, G. (1983). *Introduction to Survey Sampling.* Beverly Hills, Californie, Sage Publications. Kish, L. (1965). *Survey Sampling.* New York, Wiley.

Ligue des États arabes (1990). *Sampling Manual, Arab Maternal and Child Health Survey,* Basic Documentation 5. Le Caire, Projet panarabe pour le développement de l'enfant (PAPCHILD).

Macro International Inc. (1996). *Sampling Manual.* DHS-III Basic Documentation, n° 6. Calverton, Maryland.

Organisation des Nations Unies, Division de la statistique (1984). *Manuel d'enquêtes sur les ménages,* édition révisée, Études méthodologiques, n° 31. Numéro de vente : F.83.XVII.13.

_____ (1986). *Programmes de mise en place de dispositifs nationaux d'enquêtes sur les ménages* : *Sampling frames and sample designs for integrated household survey programmes,* version préliminaire. DP/UN/INT-84-014/5E. New York, Organisation des Nations Unies, Département de la coopération technique pour le développement et Bureau de statistique.

Pettersson, Hans (2001). Mission report: recommendations regarding design of master sample for household surveys of Viet Nam. Non publié. General Statistical Office. Hanoi, 25 novembre

_____ (2005). *Conception de cadres directeurs d'échantillonnage et d'échantillons-maîtres pour les enquêtes sur le ménages dans les pays en développement.* Dans *Enquêtes sur les ménages dans les pays en développement et dans les pays en transition*. Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Turner, A. (1998). Mission report to the Kingdom of Cambodia, National Institute of Statistics, 11-24 novembre. National Institute of Statistics. Phnom Penh, non publié.

_____ (1999). Mission report to United Arab Emirates, Ministry of Health and Central Department of Statistics, 23 janvier-3 février. Abou Dhabi, non publié. Central Department of Statistics.

_____ (2000). Mission report to Mozambique, Instituto Nacional de Estatistica, 13-26 août. Instituto Nacional de Estatistica. Maputo, non publié.

United States Bureau of the Census (1978). *Current Population Survey Design and Methodology.* Technical Paper 40. Washington, Bureau of the Census.

Verma, Vijay (1991). *Sampling Methods*. Training Handbook. Tokyo, Statistical Institute for Asia and the Pacific.

Chapitre 5

Documentation et évaluation de la conception des échantillons

5.1. Introduction

1. Le présent chapitre, bien que relativement bref, occupe cependant une place centrale dans le guide. La documentation et l'évaluation de la conception des échantillons en particulier et des méthodes d'enquête en général sont des questions trop fréquemment négligées dans la hâte de publier les résultats d'une enquête. Tel est particulièrement le cas des pays qui n'ont guère l'expérience des enquêtes sur les ménages, où les métadonnées sont fréquemment mal documentées dans les tableaux et les rapports. Dans certains pays, il n'est guère accordé d'attention à la documentation des procédures d'enquête ni à l'utilisation des échantillons, de sorte qu'il n'est habituellement guère demandé d'informations à ce sujet. Les chercheurs, lorsqu'ils analysent les données rassemblées lors des enquêtes, devraient également chercher à savoir quelles ont été les méthodes utilisées pour concevoir l'échantillon. Plusieurs documents devront par conséquent être établis pour exposer en détail les procédures suivies lors de l'enquête.

2. L'évaluation des résultats de l'enquête est souvent une indication qui est totalement ignorée, de sorte que des erreurs se glissent dans l'analyse. Cela est généralement dû au fait que, par suite de contraintes budgétaires, il est fréquemment impossible d'entreprendre des études et de mettre au point des méthodes formelles pour évaluer les abondantes erreurs autres que d'échantillonnage qui peuvent apparaître dans les enquêtes sur les ménages. Il y a cependant d'autres baromètres de la qualité des données qui peuvent être utilisés facilement (comme le taux de non-réponse), mais ils sont eux aussi, trop souvent, passés sous silence dans les rapports d'enquête.

3. Le présent chapitre insiste également sur l'importance qu'il y a à fournir aux usagers des informations pertinentes sur les limitations connues des données, même lorsqu'il n'a pas été réalisé d'études d'évaluation en bonne et due forme; il importe cependant de noter à cet égard que la discussion des méthodes à suivre pour évaluer les méthodes d'enquête, et elles sont nombreuses, sort du champ du présent guide[1]. Le chapitre ci-après met plutôt l'accent sur les informations à fournir aux usagers pour les aider à évaluer la qualité de l'enquête en centrant leur attention sur les aspects liés à l'échantillonnage.

1 Les études spéciales visant à évaluer des types spécifiques d'erreurs autres que d'échantillonnage dans les enquêtes sont notamment des études de suivi (il est procédé à de nouvelles entrevues pour évaluer la variabilité des réponses), les enquêtes post-énumération (pour évaluer la couverture et le contenu de l'enquête), l'utilisation d'échantillons à interpénétration (pour évaluer la variabilité due à l'enquêteur), la contre-vérification des données (pour évaluer les erreurs de mémoire des déclarants), etc.

5.2. Nécessité et types de documentation et d'évaluation des échantillons

4. Deux types de documentation peuvent être nécessaires dans les enquêtes sur les ménages. Il faut d'abord tenir un registre soigneux des procédures d'enquête et d'échantillonnage lorsqu'elles sont appliquées dans la pratique aux fins de l'enquête. En l'absence de cette documentation, des erreurs se glissent inévitablement dans l'analyse de l'enquête. Il se peut par exemple que les probabilités de sélection ne soient pas pleinement connues lors de l'analyse s'il n'a pas été tenu de registre scrupuleux.

5. Le technicien doit par conséquent faire le nécessaire pour documenter soigneusement non seulement le plan d'échantillonnage utilisé pour l'enquête réalisée mais aussi son application. La conception de l'échantillon doit fréquemment être adaptée à différentes étapes du travail sur le terrain par suite de situations imprévues. Il importe de consigner, étape par étape, toutes les procédures suivies pour mettre en œuvre le plan d'échantillonnage pour qu'il soit appliqué comme prévu. Lorsque tel n'est pas le cas, il est encore plus important de documenter toutes les dérogations à la conception initiale, même les plus mineures. Cette information sera nécessaire plus tard, au stade de l'analyse, s'il faut introduire des ajustements. En outre, ce type de documentation est indispensable pour la planification d'enquêtes futures.

6. Le second type de documentation est constitué par les rapports. Pour toute enquête sur les ménages, il convient d'établir deux types de rapports techniques : une description assez succincte et conviviale de la méthodologie suivie pour l'enquête, y compris le plan d'échantillonnage et son application, et une description plus détaillée de cette méthodologie. Dans le premier cas, il s'agira habituellement des sections « techniques » (ou des appendices) des différents rapports publiés pour analyser et interpréter les résultats de l'enquête[2], et il faudra généralement prévoir une section donnant des informations sur les limitations des données (voir ci-après).

7. Le second type de rapports techniques s'adressent surtout aux chercheurs, aux spécialistes des sciences sociales et aux statisticiens plutôt qu'aux décideurs ou au grand public; ils doivent par conséquent contenir une description plus détaillée de la méthodologie suivie et doivent constituer des documents indépendants plutôt que de faire partie d'une série de rapports techniques. Le Bureau of the Census des États-Unis (1974) a publié une excellente version d'un tel rapport. Il est évidemment préférable d'évaluer simultanément le rapport technique détaillé et les rapports d'enquête usuels, bien que le premier soit généralement établi beaucoup plus tard, si tant est qu'il en soit produit un. Il est bon aussi de publier le rapport technique, ou un résumé du rapport, dans une revue statistique afin d'en assurer la longévité.

8. Les rapports de l'un et l'autre types sont si importants qu'il est recommandé aux offices nationaux de statistiques de charger un statisticien ou un bureau spécial de les établir.

5.3. Dénomination des variables de conception

9. La présente section et les sections 5.2 à 5.7 se rapportent à la documentation du premier type, c'est-à-dire aux registres qui doivent être tenus des processus d'échantillonnage.

10. Une dénomination claire et spécifique doit être attribuée aux unités de sélection identifiées à chaque phase. Dans une conception à phases multiples, cela signifie qu'il faudra prévoir des codes

2 Des indications sur le contenu du rapport sur les conclusions d'une enquête par sondage figurent dans un rapport de la Sous-Commission sur l'échantillonnage statistique de l'ONU (Organisation des Nations Unies, 1964).

pour les unités d'échantillonnage primaires, secondaires, tertiaires et ultimes (selon le nombre de phases que comporte la conception). Normalement, un code à quatre chiffres suffira pour la première phase de la sélection et un code à trois chiffres pour les phases suivantes. Des dénominations appropriées doivent être attribuées aussi aux domaines géographiques. En outre, il faut indiquer les codes administratifs qui identifient les structures géographiques et administratives des zones auxquelles appartiendront les unités d'échantillonnage. Les unités d'analyse devront elles aussi être expressément identifiées.

Exemple

Supposons qu'un échantillon de 1 200 UPE, définies comme étant des zones d'énumération du recensement, soient sélectionnées pour une conception en deux phases, 600 UPE de chacun des deux domaines étant définies comme unités urbaines et unités rurales. Un moyen commode de coder les UPE est d'utiliser un système de numérotation allant de 0001 à 1 200. De plus, il est bon d'assigner ces codes dans le même ordre que celui qui est utilisé pour sélectionner les UPE. Cela peut être nécessaire pour pouvoir calculer les variances de l'échantillonnage. Ainsi, si les UPE rurales ont été sélectionnées en premier, elles seraient codées de 0001 à 0600, et les UPE urbaines de 0601 à 1 200. De tels systèmes de codage présentent deux avantages : premièrement, chaque UPE porte un numéro qui lui est propre et, deuxièmement, les analystes peuvent voir immédiatement, en se référant au code d'identification, si une UPE est urbaine ou rurale. À la seconde phase de l'échantillonnage, chaque UPE est porté sur la liste et il est sélectionné 20 ménages pour les entrevues. Lors de cette phase, tous les ménages figurant sur la liste devront recevoir un code à trois chiffres (ou à quatre chiffres si certaines zones d'énumération comportent plus de 999 ménages), dans ce cas également dans l'ordre de la liste. Enfin, l'on assigne les codes administratifs, selon que de besoin. Ainsi, le code 09 003 008 0128 pour le ménage 080 de l'échantillon identifierait celui-ci comme étant le quatre-vingtième ménage de la liste (et sélectionné pour une entrevue) de l'unité primaire d'échantillonnage 0128 appartenant à la circonscription civile 008 du district 003 de la province 09. De plus, le numéro de l'UPE permet de déterminer immédiatement que le ménage appartient au domaine rural. Si des informations ont été rassemblées sur les membres du ménage, chacun de ces derniers recevrait également un code unique à deux chiffres, allant de 01 à 99.

11. Une dénomination appropriée est essentielle, tout d'abord, du point de vue du contrôle de la qualité : les instructions données aux enquêteurs et les questionnaires revenant du terrain peuvent être vérifiés en se référant à une liste-cadre pour s'assurer que tous les ménages faisant partie de l'échantillon ont été interrogés. Deuxièmement, le système de codage est d'une énorme utilité pour le personnel chargé du traitement des données car il permet de procéder à des tabulations par emplacement géographique.

12. Pour les pays qui mènent simultanément plusieurs programmes d'enquête, il importe que les variables soient identifiées de façon uniforme et cohérente pour toutes les enquêtes. Ainsi, en éliminant les confusions qui peuvent surgir dans l'esprit aussi bien des producteurs que des usagers des données, l'on peut faciliter le traitement des informations et la présentation des résultats.

13. En ce qui concerne ce dernier point, il est conseillé, dans le cadre d'un programme comportant de multiples enquêtes, d'assigner des codes à tout l'univers d'UPE, comme décrit dans l'exemple précédent, et pas seulement aux UPE faisant partie de l'échantillon. En effet, les différentes enquêtes

portent fréquemment sur des UPE différentes, comme cela est souvent le cas lorsque l'on utilise des échantillons-maîtres.

14. D'une manière générale, il est encore plus nécessaire de coder les échantillons-maîtres que les échantillons qui ne doivent être utilisés que pour une enquête ponctuelle. L'une des principales utilisations des échantillons-maîtres, comme on l'a vu, est de permettre de procéder à plusieurs séries d'une même enquête. Un codage approprié des variables identifiant les phases de la sélection revêt par conséquent une importance capitale pour pouvoir suivre les parties de l'échantillon qui se chevauchent d'une enquête à l'autre. Souvent, les panels (sous-séries systématiques de l'échantillon proprement dit) utilisés par roulement sont codés aussi afin de faciliter l'identification des unités (ménages, grappes ou UPE) devant être remplacées lors des enquêtes ultérieures. Il va de soi qu'il faut leur assigner des codes d'identification propres. De plus, les ménages qui sont ajoutés à l'échantillon-maître lors des mises à jour périodiques doivent être codés comme il convient. Le système de codage doit être conçu de manière à pouvoir distinguer les anciens ménages des nouveaux.

5.4. Probabilités de sélection

15. Une information qui est souvent négligée, dans la documentation de l'échantillon, concerne les probabilités de sélection aux différentes phases. Lorsque des informations sur ce point existent effectivement, il arrive souvent qu'elles ne portent que sur la pondération de l'échantillon globale (à partir de laquelle l'on peut aisément calculer la probabilité globale) pour chaque échantillon.

16. Il faut, si l'on veut que la documentation soit appropriée, prendre en considération un détail particulièrement important lorsqu'il est procédé à un morcellement sur le terrain pendant la collecte des données, ce qui peut être le cas lorsque la taille d'un segment ou d'une grappe est trop grande. Par exemple, comme on l'a dit, il se peut qu'une grappe de taille imprévue doive être morcelée en, par exemple, quatre parties de tailles à peu près égales. Il est alors sélectionné au hasard une de ces parties en vue de l'établissement d'une liste et des entrevues. En pareil cas, la probabilité globale du segment pris comme échantillon (et des ménages et personnes qu'il comporte) est le quart de celle de la grappe initiale, de sorte que la pondération est l'inverse du quart, c'est-à-dire quatre. Ce facteur de pondération doit être reflété dans les calculs lors de l'analyse des données.

17. Il peut également y avoir un morcellement lorsqu'une habitation comprend plus d'un ménage (lorsque l'unité d'énumération est le logement). Une formule, qui n'introduit pas de distorsion, consiste à interroger tous les ménages que comporte l'habitation en question : telle est fréquemment l'approche suivie lorsqu'il n'y en a que deux. Cependant, s'il y a par exemple cinq ménages alors que l'on ne s'attendait à en trouver qu'un seul, des considérations de coût peuvent obliger à n'interroger qu'un seul d'entre eux, évidemment choisi au hasard. Dans ce cas également, il est essentiel de concilier soigneusement les taux de morcellement (1/5 en l'occurrence) de manière à pouvoir calculer avec précision la probabilité de sélection du ménage intéressé et d'ajuster ainsi comme il convient la pondération (par un facteur de 5).

18. Il est utile aussi d'enregistrer les probabilités de sélection à chaque phase de la sélection, comme indiqué au début de la présente section. Par exemple, la probabilité de sélection de chaque UPE est différente dans tous les cas où l'on utilise une *probabilité proportionnelle à la taille*. Cela est vrai même si l'échantillon global est autopondéré. Si les probabilités de sélection des UPE sont ignorées ou ne sont pas enregistrées comme il convient, il peut ne pas être possible de déterminer les pondérations

globales de l'échantillon. Par exemple, il est utile de savoir ce que sont les probabilités initiales de sélection pour pouvoir déterminer avec précision les procédures de morcellement.

5.5. Taux de réponse et taux de couverture des différentes phases de la sélection de l'échantillon

19. Dans le cadre du processus d'évaluation de la réalisation de l'enquête, il est essentiel de donner aux usagers des informations sur les taux de réponse et les taux de couverture. Il convient, à cet égard, de fournir des informations aussi détaillées que possible. Il importe par conséquent d'indiquer non seulement les taux de réponse (ou son corollaire, le taux de non-réponse), mais aussi une indication des raisons des non-réponses. Les catégories de non-réponses peuvent notamment être les suivantes :

- Personne à la maison;
- Logement inoccupé;
- Logement démoli ou inhabitable;
- Refus;
- Absence temporaire (congé, etc.).

20. La définition du taux de réponse, pour ce qui est des catégories à envisager, peut varier d'un pays à l'autre. Habituellement, cependant, l'on cherche à obtenir une réponse dans les premier, quatrième et cinquième types de situations. Les logements inoccupés et démolis sont habituellement ignorés (pour calculer le taux de réponse) étant donné que, par définition, il ne sera pas possible d'obtenir de réponse. Ainsi, par exemple, un pays pourra sélectionner 5 000 ménages avec les résultats suivants : 4 772 entrevues complètes, 75 cas de « personne à la maison », 31 logements inoccupés, 17 logements démolis, 12 refus et 93 cas d'« absence temporaire ». Le taux de réponse serait calculé, généralement en faisant abstraction des logements inoccupés ou démolis, comme étant 4 772/(5 000 – 31-17) soit 96,4 %.

21. Lorsque les populations cibles visées par une enquête comprennent à la fois des ménages (pour des variables comme le revenu des ménages ou l'accès aux services, et des individus) pour évaluer, par exemple, l'état de santé des femmes adultes, l'on calcule habituellement les taux de réponse aux niveaux aussi bien des ménages que des individus. Il se peut par exemple que 98 % des ménages répondent mais qu'une petite proportion des individus constituant les ménages déclarants ne répondent pas.

22. Souvent, des grappes tout entières ne sont pas interrogées pour différentes raisons, dont des considérations de sécurité, notamment en présence de troubles civils, ou des difficultés d'accès dues à la topographie de la localité ou au mauvais temps. Fréquemment, lorsque surgissent de tels problèmes, il est sélectionné des grappes de remplacement, procédure qui risque d'introduire de graves distorsions car les habitants des grappes prises comme remplacement présentent presque toujours des caractéristiques très différentes de celles des grappes remplacées. Néanmoins, lorsqu'il est procédé à de telles substitutions, les enquêteurs doivent enregistrer le nombre et l'emplacement de ces grappes. Il importe aussi, en pareil cas, de fournir quelques informations sur les raisons d'une couverture incomplète, par exemple en estimant, dans la mesure du possible, le nombre de personnes faisant partie de la population cible qui habitent sans doute dans les grappes remplacées.

23. Il convient de noter que les problèmes de ce type peuvent être quelque peu atténués si l'on identifie avant la sélection de l'échantillon les zones du pays qui seront exclues de l'enquête pour des

raisons de sécurité et d'accessibilité. Les zones ainsi identifiées devront être documentées et exclues de l'univers de l'enquête avant la sélection de l'échantillon. Les zones exclues devront être indiquées dans le rapport, et il devra être précisé que les résultats de l'enquête ne leur sont pas applicables.

5.6. Pondération : pondérations de base, non-réponses et autres ajustements

24. Le calcul des pondérations est une question évoquée au chapitre 6. Le présent chapitre met l'accent sur la nécessité de documenter ces calculs.

25. La pondération, dans le cas des enquêtes sur les ménages, comporte généralement jusqu'à trois opérations : calcul des pondérations de base ou de conception, ajustements pour tenir compte des non-réponses, et ajustements reflétant la post-stratification. Souvent, l'on n'utilise que les pondérations de conception mais, dans d'autres cas, celles-ci peuvent être ajustées par un facteur supplémentaire pour refléter les non-réponses. Dans des cas relativement rares, la pondération peut refléter un autre facteur, avec ou sans ajustement pour les non-réponses, visant à ajuster la répartition de la population faisant partie de l'échantillon de manière qu'elle corresponde à la distribution ressortant d'une source indépendante de données, comme un recensement récent. C'est ce que l'on appelle souvent une pondération post-stratification. Dans certains cas, il n'est procédé à aucune pondération, mais seulement lorsque deux conditions sont réunies : l'échantillon doit être totalement autopondéré et les données générées doivent se présenter simplement sous forme de répartitions en pourcentage, de proportions et de ratios, par opposition à des estimations ou à des chiffres absolus.

26. Lorsqu'il est procédé à une pondération, il faut évidemment consigner soigneusement les calculs effectués. Comme on l'a vu, les pondérations (ou probabilités) à chaque phase de la sélection doivent être calculées et enregistrées. En outre, il y a lieu de consigner les différentes pondérations attribuées à chaque phase de la collecte et du traitement des données, c'est-à-dire : *a*) les pondérations de conception; *b*) les pondérations de conception après multiplication par les facteurs d'ajustement pour non-réponses; et *c*) les pondérations de conception envisagées sous *b* après ajustement de post-stratification.

27. Il importe de noter que les pondérations de conception varieront d'un domaine à un autre dans tous les cas où la conception de l'échantillon comporte une estimation des domaines. Autrement dit, même lorsque l'échantillon est autopondéré à l'intérieur d'un même domaine, chaque domaine aura sa propre pondération. En outre, chaque domaine sera caractérisé par une série différente de pondérations si la conception n'est pas autopondérée à l'intérieur du domaine considéré. En outre, il y a lieu de noter que les ajustements pour non-réponses sont souvent appliqués séparément par grandes sous-régions géographiques, sans égard à la question de savoir si la conception prévoit une estimation des domaines. Enfin, la pondération de conception elle-même peut être multipliée par un facteur supplémentaire pour des grappes ou des ménages déterminés, en particulier lorsqu'il est procédé à un morcellement (voir la section 5.3).

5.7. Informations concernant les coûts de l'échantillonnage et de la réalisation de l'enquête

28. Le budget des enquêtes sur les ménages étant habituellement établi très soigneusement, il importe de comptabiliser les dépenses effectivement encourues pour les différentes opérations. Cela est

utile, par exemple, afin de planifier les opérations d'établissement des échantillons-maîtres et la tenue des registres, ce qui facilitera l'organisation des enquêtes suivantes.

29. Lorsque l'on utilise des échantillons-maîtres, il faut généralement prévoir un important investissement initial. En effet, il faut généralement : *a*) manipuler sur ordinateur les fichiers du recensement pour établir le cadre d'échantillonnage; *b*) préparer des cartes pour identifier les unités primaires d'échantillonnage; et *c*) sélectionner par ordinateur les UPE devant faire partie de l'échantillon. Comme mentionné dans le chapitre précédent, l'investissement initial est souvent partagé entre les ministères qui utiliseront l'échantillon-maître pendant tout son cycle de vie. Ces coûts devront également être répartis entre toutes les enquêtes pour lesquelles l'échantillon-maître doit être utilisé dans la mesure où elles sont connues à l'avance. Il est donc essentiel de tenir des états très rigoureux des dépenses encourues lors de l'élaboration de l'échantillon-maître ainsi que de la planification des opérations d'échantillonnage concernant de futures enquêtes.

30. Une fois l'échantillon-maître en place, il faudra tenir un registre des dépenses encourues pour sa maintenance. Comme indiqué ci-dessus, les échantillons-maîtres doivent être mis à jour périodiquement, et les dépenses correspondantes devront évidemment être suivies de près.

31. Les opérations d'échantillonnage dont le coût doit être calculé régulièrement sont notamment celles qui figurent sur la liste ci-après, qui vaut aussi bien pour les enquêtes ponctuelles que pour l'établissement d'échantillons-maîtres :

a) Traitements du personnel chargé de la conception de l'échantillon, y compris, le cas échéant, les honoraires de consultants de l'extérieur;

b) Dépenses encourues sur le terrain pour la mise à jour du cadre d'échantillonnage, y compris les traitements du personnel et l'établissement de documents auxiliaires comme des cartes;

c) Coûts des services informatiques liés à la préparation du cadre devant être utilisé pour la sélection des UPE;

d) Dépenses afférentes au personnel chargé de sélectionner l'échantillon d'UPE (si ce travail n'est pas fait par ordinateur);

e) Dépenses encourues sur le terrain pour l'établissement, à l'avant-dernière phase de l'échantillonnage, de listes d'UPE, y compris dépenses de personnes et dépenses afférentes à la préparation de fichiers de grappes;

f) Rémunération du personnel chargé de rassembler les données concernant les ménages faisant partie des grappes prises comme échantillon.

32. Indépendamment des dépenses afférentes à l'échantillonnage, il faudra également tenir un registre des coûts de réalisation de l'enquête : rémunération des enquêteurs et des superviseurs; indemnité journalière de subsistance et faux frais du personnel permanent de l'organisation chargé de l'enquête; frais de voyage; fournitures de bureau; dépenses de formation; dépenses de carburant; services de communication; et traitement des données.

5.8. Évaluation : limitations des données provenant de l'enquête

33. Pour une large part, les documents et les registres susmentionnés non seulement seront importants pour le traitement des résultats de l'enquête mais seront aussi utiles pour évaluer différents aspects de la conception de l'échantillonnage et de la réalisation de l'enquête. Les informations tou-

chant les taux de réponse, par exemple, aideront à déterminer si le biais imputable aux non-réponses est grave ou non. Les informations concernant le coût des opérations d'échantillonnage pourront servir à évaluer l'efficacité « économique » de la conception de l'échantillon et son utilité pour de futures enquêtes.

34. Comme on l'a vu, l'évaluation formelle des enquêtes par sondage porte sur de multiples aspects des erreurs autres que d'échantillonnage qui dépassent de beaucoup le champ du présent guide [voir Organisation des Nations Unies (1984) pour une étude détaillée de cette question]. Il importe de mentionner cependant que l'évaluation des erreurs autres que d'échantillonnage doit, par exemple, porter notamment sur les aspects opérationnels et le traitement des données. D'un autre côté, l'on peut estimer les erreurs d'échantillonnage, comme on le verra ci-après.

35. Bien qu'il soit rare que des études d'évaluation formelles soient entreprises au sujet des enquêtes sur les ménages, il importe au plus haut point que la documentation comprennent des informations sur les limitations des données. Les rapports sur les constatations retirées devront donc consacrer une brève section à ce sujet, qui pourra être intitulée « Limitations des données provenant de l'enquête ». Dans cette section, le lecteur devra être informé des erreurs dues à l'échantillonnage et à d'autres facteurs que comporte l'enquête.

36. Une utile publication du Bureau of the Census des États-Unis (1974) décrit comment doit être présentée l'information concernant les erreurs qui caractérisent les enquêtes. Les paragraphes ci-après, tirés de cette publication (appendice I, p. I-1), suggèrent le type d'information qu'il y aura lieu de fournir aux usagers lors de la publication des résultats de l'enquête :

> Les statistiques figurant dans le présent rapport sont des estimations tirées d'une enquête par sondage. Les estimations provenant d'une enquête par sondage peuvent comporter deux types d'erreurs : d'échantillonnage et autres que d'échantillonnage. Les erreurs d'échantillonnage sont dues au fait que les observations portent uniquement sur un échantillon et non sur l'ensemble de la population. Les erreurs autres que d'échantillonnage (dont il est question au chapitre 8) sont imputables à de nombreuses raisons : impossibilité d'obtenir des informations sur tous les cas de l'échantillon, problèmes de définition, différences d'interprétation des questions, impossibilité ou refus des déclarants de fournir des informations exactes, erreurs d'enregistrement ou de codage des données obtenues et autres erreurs concernant la collecte, les réponses, les traitements, la couverture et l'estimation des données manquantes. Il y a également des erreurs autres que d'échantillonnage lors des recensements. L'exactitude des résultats d'une enquête dépend des effets conjugués des erreurs d'échantillonnage et autres que d'échantillonnage.
>
> L'échantillon utilisé aux fins de la présente enquête est l'un des plus nombreux de tous ceux qui auraient pu être sélectionnés pour un échantillon de même conception. Les estimations tirées des différents échantillons varieraient. L'écart entre l'estimation concernant un échantillon et la moyenne de tous les échantillons possibles est appelé erreur d'échantillonnage. L'erreur type d'une estimation mesure l'écart entre les estimations et les échantillons possibles et reflète par conséquent la précision d'une estimation, par rapport au résultat moyen de tous les échantillons possibles. L'erreur type relative est définie comme étant l'erreur type divisée par la valeur estimée.
>
> Telle qu'elle a été calculée aux fins du présent rapport, l'erreur type mesure en partie aussi l'effet des erreurs autres que d'échantillonnage mais pas les biais systématiques, le cas échéant, des

données. Le biais est la différence, exprimée sous forme de moyenne pour tous les échantillons possibles, entre l'estimation et la valeur souhaitée. Manifestement, l'exactitude des résultats d'une enquête dépend des erreurs aussi bien d'échantillonnage que des autres, mesurées par l'erreur type, et le biais et les autres types d'erreurs autres que d'échantillonnage qui ne sont pas mesurés par l'erreur type.

37. Comme l'implique ce qui précède, un aspect important de l'évaluation de l'échantillon est l'estimation des erreurs d'échantillonnage, laquelle doit être entreprise pour les principales estimations. Comme on l'a dit, l'une des caractéristiques distinctives d'un échantillon probabiliste est que l'échantillon lui-même peut être utilisé pour estimer les erreurs types. Les méthodes d'estimation de la variance et de l'erreur type sont discutées en détail dans le chapitre 6. En outre, il existe des logiciels efficaces et fiables qui permettent d'estimer les erreurs types, et il y aura lieu de les utiliser dans tous les cas où cela sera possible.

38. De manière générale, il est établi des estimations des erreurs types pour les principales caractéristiques visées par l'enquête étant donné qu'il n'est ni réaliste, ni nécessaire, de les calculer pour toutes les variables. Les erreurs types constituent pour les usagers le moyen d'évaluer la fiabilité des estimations et de définir les intervalles de confiance de part et d'autre des estimations.

39. Les erreurs types peuvent également être utilisées pour évaluer la conception de l'échantillon lui-même. Une statistique particulièrement utile, à cette fin, est l'effet de conception de l'échantillon, le deff, ou plus précisément le *deft,* qui est la racine carrée du deff. Il est assez simple de calculer le *deft* pour chacune des données à propos de laquelle il y a lieu d'estimer l'erreur type. Il suffit de diviser l'erreur type estimative concernant une variable donnée par l'erreur type caractérisant un échantillon aléatoire simple de même taille, à savoir pq/n, où p est la proportion estimative; q est $1 - p$ et n est la taille de l'échantillon. Ce calcul permet de confirmer ou de réfuter les effets de conception pris comme hypothèse lors de la conception de l'échantillon étant donné que le *deff* ou le *deft* effectif ne peut être connu qu'après la réalisation de l'enquête, lorsque les données ont été traitées et que les erreurs types ont été estimées.

40. Le statisticien peut se fonder sur les effets de conception calculés pour déterminer si les grappes sont de taille raisonnable pour les principales variables et introduire les mesures correctives appropriées, le cas échéant. Par exemple, si le *deft* est beaucoup plus important que prévu pour certaines variables clés, l'échantillon destiné à une enquête future pourra être conçu de manière à utiliser des grappes de plus petites tailles.

5.9. Résumé des lignes directrices à suivre

41. La présente section résume les principales lignes directrices pouvant être tirées de ce chapitre. Comme dans le cas du chapitre 3, il s'agit d'indications plutôt que de recommandations fermes. Ainsi, il convient :

- De documenter de deux façons l'aspect échantillonnage des enquêtes : tenir des registres appropriés et fournir les informations techniques voulues aux usagers.
- De tenir des états détaillés des opérations d'échantillonnage, y compris de leurs coûts.
- D'établir des codes pour les variables liées à la conception de l'échantillon : circonscriptions administratives, unités primaires d'échantillonnage, grappes, ménages, individus, etc.

- De s'efforcer d'appliquer aux variables des codes normalisés qui soient uniformes d'une enquête à l'autre.
- D'enregistrer tous les écarts par rapport au plan d'échantillonnage initial qui se produisent pendant la réalisation de l'enquête.
- De calculer et enregistrer les probabilités de sélection à chaque phase de l'échantillonnage.
- De consigner en particulier les informations concernant le morcellement opéré pendant les travaux sur le terrain.
- De consigner les informations concernant le nombre et les types de non-réponses.
- De consigner les pondérations utilisées et les ajustements pour non-réponses et ajustements post-stratification.
- De tenir des états détaillés du coût de chaque opération de conception et d'application de l'échantillon.
- Lorsqu'il est établi un échantillon-maître, de tenir un registre des coûts d'élaboration et des coûts de maintenance.
- D'établir à l'intention des usagers des rapports techniques concernant les méthodes d'échantillonnage et d'enquête.
- D'inclure une brève section concernant les limitations des données dans toutes les publications qui rendent compte des résultats de l'enquête.
- D'élaborer un rapport technique plus détaillé sur tous les aspects de la méthodologie d'échantillonnage.
- De calculer les erreurs d'échantillonnage pour les variables clés et les indiquer dans les rapports techniques.
- De calculer les effets de conception (*deff* ou *deft*) pour les principales variables.
- De désigner une personne qui sera responsable de la documentation.

Références et autres lectures

Banque mondiale (1999). *Core Welfare Indicators Questionnaire (CWIQ) Handbook.* Washington, D.C., chapitre 4.

Casley, D. J. et D. A. Lury (1981). *Data Collection in Developing Countries.* Oxford, United Kingdom, Clarendon Press.

International Statistical Institute (1975). *Manual on Sample Design*. World Fertility Survey Basic Documentation, Voorburg (Pays-Bas).

Ligue des États arabes (1990). *Sampling Manual, Arab Maternal and Child Health Survey,* Basic Documentation 5. Le Caire. Projet panarabe pour le développement de l'enfant (PAPCHILD).

Macro International, Inc. (1996). *Sampling Manual.* DHS-III Basic Documentation n° 6. Calverton, Maryland.

Organisation des Nations Unies (1964). Recommendations for the Preparation of Sample Survey Reports (Provisional Issue). Statistical Papers, Series C, n° 1, Rev.2.

_____ (1984). *Manuel d'enquêtes sur les ménages*, édition révisée, Études méthodologiques, n° 31. Numéro de vente : F.83.XVII.13.

United States Bureau of the Census (1974). *Standards for Discussion and Presentation of Errors in Data.* Technical Paper 32. Washington, Bureau of the Census.

_____ (1978). *Current Population Survey Design and Methodology.* Technical Paper 40. Washington, Bureau of the Census.

Chapitre 6

Construction et utilisation des pondérations d'échantillonnage

6.1. Introduction

1. Le présent chapitre analyse différentes étapes de la définition des pondérations d'échantillonnage ainsi que l'utilisation qui en est faite pour calculer sur la base des données recueillies lors des enquêtes sur les ménages des estimations des caractéristiques à étudier. En particulier, il décrit l'ajustement à apporter aux pondérations pour compenser les différentes imperfections de l'échantillon sélectionné. La discussion porte uniquement sur les estimations descriptives qui constituent les produits les plus fréquents de la plupart des enquêtes. Les idées importantes qui y sont présentées sont illustrées au moyen d'exemples réels d'enquêtes menées dans des pays en développement ou d'enquêtes simulant des situations réelles.

6.2. Nécessité de pondérer les échantillons

2. D'une manière générale, les enquêtes sur les ménages sont fondées sur des échantillons de conception complexe, essentiellement afin de maîtriser les coûts. Les échantillons ainsi élaborés présentent généralement des imperfections qui peuvent se traduire par des biais et d'autres écarts entre l'échantillon et la population de référence. De telles imperfections peuvent tenir notamment à la sélection d'unités de probabilités inégales, à la non-couverture de la population cible et au phénomène de non-réponse. Les échantillons doivent être pondérés pour corriger ces imperfections et ainsi pouvoir tirer des estimations appropriées des caractéristiques visées. En résumé, la pondération a pour but :

a) De compenser des probabilités inégales de sélection;

b) De compenser le défaut de réponse des unités d'échantillonnage;

c) D'ajuster la répartition de l'échantillon pondéré en fonction des principales variables à étudier (par exemple âge, race et sexe) de sorte qu'elles correspondent à la répartition d'une population connue.

3. Les procédures suivies pour chacun des scénarios sont examinées en détail dans les sections ci-après. Une fois que les imperfections de l'échantillon ont été compensées, les pondérations peuvent être utilisées pour estimer les caractéristiques à identifier ainsi que pour estimer les erreurs d'échantillonnage pouvant affecter les estimations produites.

4. Lorsque des pondérations ne sont pas utilisées pour compenser des taux différenciés de sélection à l'intérieur des strates (lorsque l'échantillon est ainsi conçu) et les imperfections susmentionnées,

les estimations concernant la population en résultant refléteront généralement un biais. (Voir les sections 6.3, 6.4 et 6.5 pour des exemples de procédures de pondération employées dans le cadre de chaque scénario, y compris, dans chaque cas, une comparaison des estimations pondérées et non pondérées.)

6.2.1. Aperçu général

5. La section 6.3 a trait à l'identification des pondérations à affecter dans le contexte d'une conception en plusieurs étapes, y compris pour ce qui est des ajustements à apporter aux pondérations pour tenir compte des doubles emplois que comporte l'échantillon et des unités dont on ne sait pas, lors de la sélection de l'échantillon, si elles doivent être incluses dans l'enquête. La section 6.4 traite des pondérations à utiliser pour compenser des probabilités inégales de sélection elle contient plusieurs exemples numériques, d'après une étude de cas, de l'identification des pondérations à utiliser pour une enquête nationale sur les ménages et, en conclusion, évoque les échantillons autopondérés. Les questions liées à la non-réponse et à la non-couverture sont examinées aux sections 6.5 et 6.6, respectivement, qui discutent des causes et des conséquences de ces phénomènes. Ces sections exposent également les méthodes qui peuvent être suivies pour compenser les non-réponses et la non-couverture et donnent notamment des exemples numériques illustrant l'effet de l'ajustement apporté aux pondérations de l'échantillon. La section 6.7 examine la question de l'inflation dans le contexte de la variation des estimations issues de l'enquête entraînée par l'utilisation de pondérations. Cette section donne également un exemple numérique pour illustrer les calculs de l'accroissement de la variation dû à la pondération. La section 6.8 aborde la question de l'allégement des pondérations et donne un exemple de procédure d'allégement selon laquelle les pondérations allégées sont réaménagées de sorte qu'elles correspondent ensemble à la somme des pondérations initiales. La section 6.9, enfin, offre quelques conclusions.

6.3. Identification des pondérations d'échantillonnage

6. Une fois que les probabilités de sélection des unités devant faire partie de l'échantillon ont été déterminées, l'on peut commencer à identifier les pondérations d'échantillonnage. La probabilité de sélection d'une unité dépend de la conception utilisée pour la sélectionner. L'on a vu au chapitre 3 quelles étaient les conceptions d'échantillon les plus communément utilisées et les probabilités de sélection correspondant à chacune d'elles. Il a été pris pour hypothèse, dans tous les cas, que les probabilités de sélection ont été déterminées.

7. L'identification des pondérations d'échantillonnage est parfois considérée comme étant la première étape des analyses des données recueillies lors de l'enquête. L'on commence habituellement à construire la *pondération de base* pour chaque unité faisant partie de l'échantillon de manière à refléter leurs probabilités inégales de sélection. La pondération de base d'une unité prise comme échantillon est la réciproque de sa probabilité de sélection ou d'inclusion dans l'échantillon. Selon sa représentation mathématique, si une unité est incluse dans l'échantillon avec une probabilité p_i, sa pondération de base, c'est-à-dire w_i, est égale à :

$$w_i = 1/p_i \qquad (6.1)$$

8. Par exemple, une unité sélectionnée sur la base d'une probabilité égale à 1/50 représente 50 unités de la population dont l'échantillon a été tiré. Ainsi, les pondérations d'échantillonnage font fonc-

tion de facteurs d'inflation conçus de manière à représenter le nombre d'unités de la population qui sont représentées par l'unité d'échantillonnage à laquelle la pondération est affectée. La somme des pondérations donne une estimation non biaisée du nombre total d'unités de la population cible.

9. Pour les conceptions à plusieurs phases, les pondérations de base doivent refléter les probabilités de sélection à chaque phase. Par exemple, dans le cas d'une conception en deux phases dans laquelle la $i^{ième}$ UPE est sélectionnée sur la base d'une probabilité p_i lors de la première phase tandis que le $j^{ième}$ ménage sélectionné sur la base d'une UPE sélectionnée sur la base d'une probabilité $p_{j(i)}$ lors de la deuxième phase, la probabilité globale de sélection (p_{ij}) de chaque ménage faisant partie de l'échantillon est donnée par le produit de ces deux probabilités, ou :

$$p_{ij} = p_i * p_{j(i)} \qquad (6.2)$$

et la pondération de base globale de ménages est obtenue comme précédemment, en prenant la réciproque de sa probabilité globale de sélection. De même, si la pondération de base du $j^{ième}$ ménage est $w_{ij,b}$, la pondération imputable à la compensation des cas de non-réponse est $w_{ij,nr}$, et la pondération imputable à la compensation de la non-couverture est $w_{ij,nc}$, de sorte que la pondération globale du ménage est donnée par la formule :

$$w_{ij} = w_{ij,b} * w_{ij,nr} * w_{ij,nc} \qquad (6.3)$$

6.3.1. Ajustements des pondérations d'échantillonnage visant à compenser les admissibilités inconnues

10. Il y a parfois, lors de la collecte de données dans le cas d'enquêtes sur les ménages, des cas dans lesquels l'admissibilité d'un ménage est douteuse. Il se peut par exemple que l'enquêteur ne trouve personne à la maison lors de l'enquête ou après des visites répétées. En pareil cas, l'on ne sait pas si le logement est occupé ou non. S'il l'est effectivement, il y a lieu de le ranger dans la catégorie des logements sans réponse (sous la rubrique « personne à la maison »). Autrement, il n'entre pas dans le champ de l'enquête, de sorte qu'il ne peut pas être dénombré comme unité d'échantillonnage. Parfois, l'enquêteur tient pour acquis que s'il ne trouve personne à la maison à la suite de visites répétées, le logement est inoccupé et par conséquent inadmissible. Cela est généralement une hypothèse inexacte qui entraîne souvent un gonflement indu des taux de réponse.

11. Lorsque l'admissibilité de certains logements est inconnue, leurs pondérations doivent être ajustées pour en tenir compte, l'idée étant de faire un certain nombre d'hypothèses qui permettent d'estimer la proportion des logements dont l'admissibilité est inconnue qui sont effectivement admissibles. La méthode la plus simple consiste à prendre la proportion de logements dont on sait qu'ils sont admissibles ou inadmissibles, et de l'appliquer à ceux dont l'admissibilité est inconnue. Supposons par exemple qu'un échantillon de 300 logements présente les taux de réponse indiqués au tableau 6.1.

12. Il y a lieu de noter que la proportion de logements dont l'admissibilité est connue qui sont effectivement admissibles est de (215 + 25)/(215 + 25 + 10) = 0,96. L'on peut par conséquent prendre pour hypothèse que 0,96 des logements dont l'admissibilité est inconnue peuvent être considérés comme admissibles. Autrement dit, 96 % des 50 logements dont l'admissibilité est inconnue (soit 48 logements) sont effectivement admissibles. Il y a lieu d'ajuster alors les pondérations des logements

admissibles (entrevues complètes et non-déclarants admissibles) sur la base d'un facteur d'ajustement défini comme suit :

$$F_{ue} = \frac{\sum_{c} w_{ij,b} + \sum_{nr} w_{ij,b} + \varepsilon \times \sum_{ue} w_{ij,b}}{\sum_{c} w_{ij,b} + \sum_{nr} w_{ij,b}} \quad (6.4)$$

où ε dénote la proportion de cas d'admissibilité inconnue dont on estime qu'ils sont admissibles (dans cet exemple, $\varepsilon = 0{,}96$). Les sommes qui figurent en exposant de *c*, *nr* et *ue* dans la formule ci-dessus dénotent respectivement la somme des pondérations de base des logements ayant donné lieu à des entrevues complètes, des non-déclarants admissibles et des logements dont l'admissibilité est inconnue. Les pondérations de base ajustées des logements ayant donné lieu à des entrevues complètes et des non-déclarants admissibles sont alors obtenues en multipliant les pondérations de base initiales $w_{ij,b}$ par le facteur F_{ue}.

Tableau 6.1
Catégories de réponses lors d'une enquête

Catégorie de réponse	Nombre de logements
Entrevues complètes	215
Non-déclarants admissibles	25
Inadmissibles	10
Admissibilité inconnue	50

6.3.2. Ajustements des pondérations d'échantillonnage pour tenir compte des doubles inscriptions sur les listes

13. Si l'on sait à priori que certaines unités sont inscrites deux fois dans le cadre d'échantillonnage, leur probabilité de sélection accrue peut être compensée en leur affectant des facteurs de pondération qui sont les réciproques du nombre de doubles inscriptions sur les listes si les unités en question sont finalement sélectionnées. Il arrive fréquemment, toutefois, que les doubles inscriptions ne soient découvertes qu'après la sélection des échantillons, de sorte que les probabilités de sélection de ces unités doivent être ajustées pour compenser les doubles inscriptions. Cet ajustement est opéré comme suit. Supposons que la $i^{ième}$ unité sélectionnée ait une probabilité de sélection dénotée par p_{i1} et que le cadre d'échantillonnage comporte $k - 1$ autres entrées, faisant double emploi, pour cette unité, chacune avec des probabilités de sélection données par p_{i2}, ..., p_{ik}. Ainsi, la probabilité de sélection ajustée de l'unité en question est donnée par :

$$p_i = 1 - (1 - p_{i1})(1 - p_{i2}) \ldots (1 - p_{ik}) \quad (6.5)$$

L'unité est alors pondérée en conséquence, c'est-à-dire par $1/p_i$.

14. Les procédures d'estimation des pondérations d'échantillonnage dans les scénarios indiqués ci-dessus sont illustrées par les exemples ci-après.

6.4. Pondérations visant à compenser les probabilités inégales de sélection

15. Pour la simplicité de l'exposé, considérons une conception en deux phases, les UPE étant les zones d'énumération du recensement et les unités secondaires les ménages. Supposons qu'il soit sélectionné un échantillon de n UPE sur la base d'une probabilité égale sur un total de N à la première phase, et que m ménages soient alors sélectionnés parmi chaque UPE faisant partie de l'échantillon. La probabilité de sélection d'un ménage dépendra manifestement du nombre total de ménages que comporte l'UPE dont il fait partie. Supposons en outre que M_i dénote le nombre de ménages que comporte l'UPE$_i$. Ainsi, la probabilité de sélection d'une UPE est de n/N et la probabilité conditionnelle de sélection d'un ménage de la $i^{ème}$ UPE est m/M_i. Par conséquent, la probabilité globale de sélection d'un ménage est donnée par la formule :

$$p_{ij} = p_i \times p_{j(i)} = \frac{n}{N} \times \frac{m}{M_i} = \frac{nm}{N} \times \frac{1}{M_i} \tag{6.6}$$

En outre, la pondération d'un ménage, selon cette conception, est donnée par

$$w_i = \frac{1}{p_{ij}} = \frac{N}{nm} \times M_i \tag{6.7}$$

Exemple 1

Il est sélectionné sur un total de 250 ménages, sur la base d'une probabilité égale, un échantillon de 5 ménages. Un adulte est sélectionné au hasard au sein de chaque ménage. Le revenu mensuel (y_{ij}) et le niveau d'instruction (z_{ij} = 1 pour le niveau secondaire et au-dessus et est autrement égal à 0) du $j^{ième}$ adulte du $i^{ième}$ ménage sont enregistrés. M_i dénote le nombre d'adultes que comporte le ménage i. Ainsi, la probabilité globale de sélection d'un adulte est donnée par :

$$p_{ij} = p_i \times p_{j(i)} = \frac{5}{250} \times \frac{1}{M_i} = \frac{1}{50} \times \frac{1}{M_i}.$$

Par conséquent, la pondération d'un adulte faisant partie de l'échantillon est donnée par :

$$w_i = \frac{1}{p_{ij}} = 50 \times M_i.$$

16. Nous allons maintenant illustrer le calcul des estimations de base dans le cas de la conception susmentionnée. Supposons que les données obtenues de l'adulte pris comme échantillon dans chaque ménage qui fait partie de l'échantillon primaire de cinq ménages soient celles indiquées au tableau 6.2 ci-après. Il y a lieu de noter que le nombre d'adultes que comporte chaque ménage et la pondération globale correspondante de l'adulte pris comme échantillon dans chaque ménage figurent dans les deuxième et troisième colonnes respectivement.

Tableau 6.2
Pondérations dans les cas de probabilités inégales de sélection

Ménage pris comme échantillon	M_i	w_i	y_{ij}	z_{ij}	$w_i y_{ij}$	$w_i z_{ij}$	$w_i z_{ij} y_{ij}$
1	3	150	70	1	10 500	150	10 500
2	1	50	30	0	1 500	0	0
3	3	150	90	1	13 500	150	13 500
4	5	250	50	1	12 500	250	12 500
5	4	200	60	0	12 000	0	0
Total	16	800	300	3	50 000	550	36 500

17. Les estimations des différentes caractéristiques peuvent alors être tirées du tableau 6.2, comme suit :

L'estimation du revenu mensuel moyen est :

$$\bar{y}_w = \frac{\sum w_i y_{ij}}{\sum w_i} = \frac{50\ 000}{800} = 62{,}5.$$

S'il n'avait pas été utilisé de pondération, cette estimation serait de 60 (ou 300/5).

L'estimation de la proportion de personnes ayant un niveau d'instruction égal ou supérieur au secondaire est :

$$\bar{y}_w = \frac{\sum w_i z_{ij}}{\sum w_i} = \frac{550}{800} = 0{,}6875 \text{ ou } 68{,}75\ \%.$$

S'il n'avait pas été utilisé de pondération, cette estimation serait de 3/5 ou 60 %.

L'estimation du nombre total de personnes ayant un niveau d'instruction égal ou supérieur au secondaire est :

$$\hat{t} = \sum w_i z_{ij} = 550.$$

L'estimation du revenu mensuel moyen des adultes ayant un niveau d'instruction égal ou supérieur au secondaire est :

$$\bar{y}_w = \frac{\sum w_i z_{ij} y_{ij}}{\sum w_i z_{ij}} = \frac{36\ 500}{550} = 66{,}36.$$

18. Parfois, les pondérations d'échantillonnage sont « normalisées », autrement dit les pondérations sont multipliés par le ratio :

$$\frac{\text{nombre des déclarants}}{\text{somme des pondérations de tous les déclarants}} \qquad (6.8)$$

19. Ainsi, la somme des pondérations normalisées est la taille réelle de l'échantillon devant servir à l'analyse (nombre de déclarants). Il y a lieu de noter que les pondérations normalisées ne peuvent pas être utilisées pour estimer des totaux, comme le nombre total d'adultes ayant un niveau d'instruction égal ou supérieur au secondaire. En pareil cas, les unités prises comme échantillon doivent être pondérées par la réciproque de leurs probabilités de sélection, c'est-à-dire qu'il faut utiliser les

pondérations d'échantillonnage ordinaire. Cependant, pour estimer des moyens et des proportions, il suffit que les pondérations soient proportionnelles aux réciproques des probabilités de sélection. Autrement dit, peu importe que des pondérations ordinaires ou des pondérations normalisées (mais qu'elles soient proportionnelles aux pondérations ordinaires) soient utilisées pour obtenir des estimations des moyennes des paramètres concernant la population comme le nombre moyen ou la proportion de femmes en âge de procréer ayant accès aux soins de santé primaires. Les deux types de pondération donneront le même résultat.

20. Par exemple, dans l'exemple précédent, les pondérations w_i' sont proportionnelles à M_i (w_i = 50 * M_i). Si nous utilisons M_i comme pondération, l'estimation de la proportion de personnes ayant un niveau d'instruction égal ou supérieur au secondaire est :

$$\hat{p} = \frac{\sum M_i z_{ij}}{\sum M_i} \frac{3\times1+1\times0+3\times1+5\times1+4\times0}{3+1+3+5+4} = \frac{11}{16} = 0{,}6875,$$

soit 68,75 % ou exactement le même résultat que précédemment. Toutefois, pour estimer le nombre total d'adultes ayant un niveau d'instruction égal ou supérieur au secondaire, il faut utiliser les pondérations ordinaires (w_i = 50 * M_i) pour obtenir le résultat correct, c'est-à-dire :

$$\hat{t}_s = \sum(50\times M_i)z_{ij} = 50\sum M_i z_{ij} = 50\times11 = 550.$$

Il est sélectionné en deux phases un échantillon de ménages dans les régions rurales dans le pays. Lors de la première phase, il est pris comme échantillon 50 villages sur la base d'une probabilité proportionnelle au nombre de ménages qu'il comportait lors du dernier recensement. Le nombre total de ménages se trouvant dans les régions rurales lors du premier recensement était de 300 000. La sélection de l'échantillon primaire a été suivie par l'établissement d'une liste des logements pour chacun des villages sélectionnés. Dans certains cas, il a été constaté que le même logement comportait plus d'un ménage.

21. Nous envisagerons maintenant différentes options pour ce qui est de la conception du sous-échantillon (pour la sélection des ménages parmi les logements sélectionnés) et spécifierons l'équation de sélection pour la probabilité globale de sélection d'un ménage. Disons que D_i dénote le nombre de logements du village i et que H_{ij} dénote le nombre de ménages du logement j du village i. Le nombre total de ménages dans un village, dénoté par H_i, est alors donné par :

$H_i = \sum_j H_{ij}$. Il y a lieu de noter que $\sum_i H_i = \sum_i \sum_j H_{ij} = 300\ 000$.

Les probabilités de sélection calculées ici sont fondées sur les formules présentées dans le chapitre 3.

Option de conception 1

22. Quinze logements sont sélectionnés sur la base d'un échantillonnage aléatoire simple sans remplacement (SRSWOR) parmi la liste établie pour chaque village sélectionné. Tous les ménages des logements sélectionnés sont inclus dans l'échantillon, de sorte qu'il n'y a que deux étapes de sélection : celle des villages et celle des logements. Selon cette conception, l'équation de sélection pour la probabilité globale de sélection d'un ménage est donné par

p_{ij} = pr (le village i est sélectionné)* pr (le logement j est sélectionné étant donné que i est sélectionné)

Ainsi :

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{15}{D_i} = \frac{750}{\sum_i M_i} \times \frac{H_i}{D_i},$$

et la pondération de base est donnée par :

$$w_{ij} = \frac{\sum_i M_i}{750} \times \frac{D_i}{H_i}.$$

23. Il y a lieu de noter que la probabilité globale de sélection variera d'un village à l'autre selon le ratio entre le nombre de ménages et le nombre de logements H_i/D_i. Par conséquent, la conclusion est que cette conception n'est pas autopondérée (pour plus amples détails sur les conceptions autopondérées, voir la section 6.4.2 ci-après). Elle serait autopondérée si chacun des logements ne comportait qu'un seul ménage, autrement dit si le ratio H_i/D_i était le même pour tous les villages de l'échantillon.

Option de conception 2

24. Les logements sont sélectionnés systématiquement dans chaque village sélectionné, le taux d'échantillonnage dans un village étant inversement proportionnel au nombre de villages qu'il comportait lors du dernier recensement. Tous les ménages des logements sélectionnés font partie de l'échantillon. Comme précédemment, la sélection comporte seulement deux étapes : celle des villages et celle des logements. La probabilité conditionnelle de sélection d'un logement d'un village sélectionné i peut être exprimée comme étant k/H_i, k étant la constante de proportionnalité. Par conséquent, le quotient de sélection pour la probabilité globale de sélection d'un ménage selon cette conception est donné par :

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} = \frac{50 \times k}{\sum_i M_i}$$

et la pondération de base est donnée par :

$$w_{ij} = \frac{\sum_i M_i}{50 \times k}$$

qui est une constante. La conclusion est par conséquent que l'option de conception 2 est une conception autopondérée.

Option de conception 3

25. Les logements sont sélectionnés systématiquement dans chaque village sélectionné, le taux d'échantillonnage dans un village étant inversement proportionnel au nombre de villages qu'il comportait lors du dernier recensement. Un ménage est sélectionné au hasard dans chaque logement sélectionné. Dans ce cas particulier, il y a trois phases de sélection : villages, logements et ménages. Par conséquent, l'équation de sélection pour la probabilité globale de sélection d'un ménage selon cette conception est donnée par :

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} \times \frac{1}{H_{ij}}$$

et la pondération de base est donnée par :

$$w_{ij} = \frac{\sum_i M_i}{50} \times \frac{H_{ij}}{k}$$

qui variera d'un logement à l'autre selon le nombre de ménages qu'il comporte. La conclusion est par conséquent que l'option de conception 3 n'est pas autopondérée.

6.4.1. Étude de cas concernant la construction de pondérations : Enquête sanitaire nationale réalisée au Viet Nam en 2001

26. Nous allons maintenant illustrer la construction des pondérations d'échantillonnage pour une enquête réelle, l'Enquête sanitaire nationale réalisée au Viet Nam en 2001. L'enquête était fondée sur une conception stratifiée en trois phases. Il y avait en tout 122 strates, définies par domaine urbain ou rural, dans 61 provinces. L'échantillon a alors été sélectionné de manière indépendante dans chaque strate. Lors de la première phase, les communes ont été sélectionnées sur la base d'une probabilité proportionnelle à la taille (nombre de ménages lors du recensement de la population et du logement de 1999). À la deuxième phase, deux zones d'énumération ont été sélectionnées dans chaque commune prise comme échantillon au moyen d'un échantillonnage systématique sur la base d'un taux d'échantillonnage inversement proportionnel au nombre de zones d'énumération de la commune. Lors de la troisième et dernière phase, 15 ménages ont été sélectionnés dans chaque zone d'énumération sélectionnée, dans ce cas également sur la base d'un échantillonnage systématique.

27. Les pondérations d'échantillonnage de base pour les ménages sélectionnés aux fins de l'Enquête sanitaire nationale peuvent être calculées comme suit. Disons que H_i et E_i dénotent respectivement le nombre de ménages et le nombre de zones d'énumération (lors du recensement de 1999) de la commune i et que H_{ij} dénote le nombre de ménages dans la zone d'énumération j de la commune i. Ainsi, la probabilité globale de sélection du ménage k dans la zone d'énumération j de la commune i est donnée par :

$$p_{ijk} = n_c \times \frac{H_i}{\sum_i H_i} \times \frac{2}{E_i} \times \frac{15}{H_{ij}}$$

où n_c est le nombre de communes sélectionnées dans une strate donnée et $\sum_i H_i$ est le nombre total de ménages que comporte la strate.

La pondération d'échantillonnage des ménages (w_{ijk}) est la réciproque de la probabilité de sélection, c'est-à-dire :

$$w_{ijk} = \frac{E_i \times H_{ij} \times \sum_i M_i}{30 \times n_c \times H_i} \qquad (6.9)$$

6.4.2. Échantillons autopondérés

28. Lorsque les pondérations de toutes les unités sélectionnées sont identiques, l'échantillon est appelé *autopondéré*. Même si, pour des raisons d'efficience, les unités primaires sont fréquemment sélectionnées selon des probabilités diverses, celles-ci se compensent par des probabilités de sélection lors des phases suivantes. Un exemple de cette situation est l'option de conception 2 de l'exemple 2 ci-dessus.

29. Dans la pratique, cependant, il est rare, pour plusieurs raisons, que les échantillons sélectionnés pour une enquête sur les ménages soient autopondérés au plan national. Premièrement, les unités d'échantillonnage sont souvent délibérément sélectionnées sur la base de probabilités inégales de sélection. En fait, bien que les UPE soient fréquemment sélectionnées sur la base des probabilités proportionnelles à la taille et que les ménages soient sélectionnés sur la base d'un taux approprié parmi les UPE de manière à donner un échantillon autopondéré, cette autopondération peut être réduite à néant par la sélection dans chaque ménage pris comme échantillon de la personne qui sera interrogée. Deuxièmement, l'échantillon sélectionné présente fréquemment des déficiences, y compris pour des raisons de non-réponse (section 6.5) et de non-couverture (section 6.6). Troisièmement, la nécessité d'obtenir des estimations précises pour certains domaines et certains sous-groupes spéciaux conduit fréquemment à sur-sélectionner ces domaines de manière à obtenir des échantillons d'une taille suffisante pour répondre aux normes de précision spécifiées. Quatrièmement, lorsque la conception de l'échantillon conduit à établir une liste à jour des ménages des grappes sélectionnées (unités primaires d'échantillonnage ou unités secondaires d'échantillonnage) et qu'un nombre fixe et prédéterminé de ménages doit être sélectionné dans chaque grappe, la probabilité effective de sélection du ménage est un peu différente de sa probabilité théorique, qui était fondée sur le dénombrement opéré lors de l'établissement du cadre plutôt que sur le nombre actuel de ménages; par conséquent, il surgit des probabilités inégales de sélection même lorsque l'on cherche à appliquer une conception autopondérée.

30. En dépit des contraintes susmentionnées, un échantillon autopondéré doit être l'idéal visé par toute conception en raison des avantages qu'il présente en ce qui concerne non seulement la mise en œuvre de la conception retenue mais aussi l'analyse des données recueillies. Lorsqu'un échantillon est autopondéré, des estimations peuvent être dérivées de données non pondérées et les résultats peuvent être « gonflés » par un facteur constant pour obtenir des estimations appropriées des paramètres de la population. De plus, les analyses fondées sur des échantillons autopondérés et les résultats sont plus faciles à comprendre et à admettre pour les non-statisticiens et le grand public.

6.5. Ajustement des pondérations d'échantillonnage pour tenir compte des non-réponses

31. Il est rare que toutes les informations souhaitées puissent, lors d'une enquête, être obtenues de toutes les unités prises comme échantillon. Il se peut par exemple que certains ménages ne communiquent pas du tout d'informations, tandis que d'autres fourniront des informations partielles seulement, c'est-à-dire répondront à certaines des questions posées mais pas à toutes. Le premier type de non-réponse est appelé *non-réponse unitaire* ou *totale,* et le second *non-réponse ponctuelle.* S'il y a des différences systématiques entre déclarants et non-déclarants, des estimations fondées exclusivement, de façon simpliste, sur les déclarants seront faussées. Une importante bonne pratique d'enquête, mise en relief dans tout le présent guide, est qu'il importe de maintenir le taux de non-réponse à un niveau

aussi bas que possible. Cela est indispensable en effet pour éviter que les estimations soient faussées d'une façon ou d'une autre en excluant (ou en n'incluant qu'une proportion excessivement réduite) un groupe de population déterminé. Par exemple, les citadins ayant un revenu relativement élevé peuvent avoir moins de chances de participer à une enquête plurivalente comprenant des questions concernant le revenu. Ne pas obtenir de réponses d'un vaste secteur de cette partie de la population pourrait affecter les estimations nationales du revenu moyen des ménages, du niveau d'instruction, du taux d'alphabétisation, etc.

6.5.1. Réduction de la distorsion due à la non-réponse dans les enquêtes sur les ménages

32. L'ampleur de la distorsion due à la non-réponse pour la moyenne d'un échantillon, par exemple, est fonction de deux facteurs :

- La proportion de la population qui ne répond pas;
- L'ampleur de la différence entre les moyennes des groupes de répondants et de non-répondants.

33. Pour réduire la distorsion due à la non-réponse, il faut par conséquent soit que le taux de non-réponse soit réduit, soit que les différences entre les ménages et personnes déclarants et non-déclarants soient réduites. En tenant un état exact de toute unité d'échantillonnage sélectionnée aux fins de l'enquête, l'on peut estimer directement, sur la base des données recueillies, le taux de non-réponse pour l'ensemble de l'échantillon et pour les sous-domaines visés. En outre, il peut être réalisé des études spéciales judicieusement conçues pour évaluer les différences entre déclarants et non-déclarants (Groves et Couper, 1998).

34. Dans le cas d'une enquête par panel (les données étant alors rassemblées à plusieurs reprises auprès des unités du même panel sur une certaine période de temps), le concepteur a accès à des données plus détaillées pour étudier et compenser les effets de la distorsion due à la non-réponse que ce n'est le cas lors d'enquêtes ponctuelles ou transversales. En l'occurrence, la non-réponse peut être due au fait que des unités d'échantillonnage se trouvent perdues lors des enquêtes de suivi ou bien refusent, par lassitude ou pour d'autres raisons, de participer aux séries d'enquêtes suivantes. Les données rassemblées lors des phases antérieures peuvent alors être utilisées pour en savoir plus sur les différences entre déclarants et non-déclarants et peuvent servir de base au type d'ajustements décrits ci-après. Les différentes techniques de compensation pour la non-réponse sont décrites par Brick et Kalton (1996) et Lepkowski (2003) et par les auteurs cités dans ces ouvrages.

6.5.2. Compensation de la non-réponse

35. Plusieurs techniques peuvent être utilisées pour accroître les taux de réponse et ainsi atténuer la distorsion due à la non-réponse et les enquêtes sur les ménages. L'une d'elles consiste à convertir une non-réponse en réponse au moyen de visites de suivi, l'enquêteur essayant non pas une mais plusieurs fois d'achever l'entrevue avec le ménage faisant partie de l'échantillon. Les taux de réponse peuvent également être améliorés lorsque les enquêteurs sont mieux formés. Cependant, quels que soient les efforts déployés pour accroître le taux de réponse, il subsistera toujours des non-réponses dans toutes les enquêtes sur les ménages. Par conséquent, les concepteurs introduisent fréquemment des ajustements pour compenser la non-réponse. Les deux approches fondamentales suivies à cette fin consistent à :

a) Ajuster la taille de l'échantillon en prenant initialement un échantillon plus nombreux que nécessaire pour tenir compte du taux prévu de non-réponse;

b) Ajuster les pondérations d'échantillonnage pour compenser la non-réponse.

36. Il est bon, dans les enquêtes sur les ménages, de toujours chercher à compenser la non-réponse ponctuelle en ajustant les pondérations d'échantillonnage pour tenir compte des ménages non-déclarants. La section 6.5.3 donne un aperçu de la démarche à suivre pour ajuster les pondérations d'échantillonnage pour compenser la non-réponse et contient un exemple numérique.

37. L'opération de remplacement, qui revient à faire une hypothèse concernant toutes les caractéristiques de l'unité non déclarante, suscite plusieurs problèmes (Kalton, 1983). Premièrement, cette méthode accroît les probabilités de sélection pour les remplaçants potentiels étant donné que les ménages non sélectionnés proches des ménages sélectionnés non répondants ont une plus forte probabilité de sélection que ceux qui sont proches de ménages sélectionnés répondants. Deuxièmement, les tentatives de remplacement de ménages non répondants prennent du temps, entraînent souvent des erreurs et des distorsions et sont très difficiles à vérifier ou à suivre. Il se peut par exemple qu'un ménage non répondant soit remplacé par un ménage commode plutôt que par le ménage pouvant véritablement constituer un remplacement, ce qui introduit une autre source de distorsion. Du fait de tous ces problèmes, il faut éviter d'avoir recours à un remplacement pour compenser la non-réponse dans les enquêtes sur les ménages, à moins qu'il n'y ait de bonnes raisons de le faire.

38. En ce qui concerne la non-réponse partielle ou ponctuelle, la méthode de compensation standard est celle de l'*imputation,* qui n'est pas traitée dans le présent guide.

6.5.3. Ajustements des pondérations d'échantillonnage pour compenser la non-réponse

39. Dans les enquêtes de grande envergure sur les ménages, l'on a fréquemment recours à la méthode consistant à ajuster les pondérations d'échantillonnage pour compenser la non-réponse. Essentiellement, l'ajustement transfère les pondérations de base de toutes les unités sélectionnées admissibles non déclarantes aux unités déclarantes. Cette technique comporte les étapes suivantes :

- *Étape 1.* Application des pondérations de conception initiales (et introduction des ajustements visant à compenser les probabilités inégales de sélection ainsi que des autres ajustements dont il est question dans les questions précédentes, s'il y a lieu).
- *Étape 2.* Décomposition de l'échantillon en sous-groupes et calcul des taux pondérés de réponse pour chaque sous-groupe.
- *Étape 3.* Utilisation de la réciproque des taux de réponse du sous-groupe aux fins des ajustements tendant à compenser la non-réponse.
- *Étape 4.* Calcul de la pondération ajustée pour non-réponse pour la $i^{ième}$ unité d'échantillonnage, selon la formule :

$$w_i = w_{1i} * w_{2i} \qquad (6.10)$$

où w_{1i} est la pondération initiale et w_{2i} la pondération tendant à compenser la non-réponse. Il y a lieu de noter que le taux pondéré de non-réponse peut être défini comme étant le ratio entre le nombre pondéré d'entrevues achevées avec des ménages sélectionnés admissibles et le nombre pondéré de ménages sélectionnés admissibles.

Exemple

Il est sélectionné en plusieurs étapes un échantillon stratifié de 1 000 ménages dans deux régions (nord et sud) du pays. Les ménages du nord sont sélectionnés sur la base d'un taux égal à 1/100 et ceux du sud de 1/200. Les taux de réponse en milieu urbain sont plus faibles qu'en milieu rural. Disons que n_h dénote le nombre de ménages sélectionnés dans la strate h, r_h le nombre de ménages admissibles qui répondent et t_h le nombre de ménages déclarants ayant accès aux soins de santé primaires. Ainsi, la pondération ajustée pour tenir compte du taux de non-réponse des ménages de la strate h est donnée par

$$w_h = w_{1h} * w_{2h} \tag{6.11}$$

où $w_{2h} = n_h/r_h$. L'on suppose que les données au niveau des strates sont celles qui sont indiquées au tableau 6.3.

Tableau 6.3
Ajustement des pondérations pour compenser la non-réponse

Strate	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h r_h$	$w_h t_h$
Nord urbain	100	80	70	100	1,25	125	10 000	8 750
Nord rural	300	120	100	100	2,50	250	30 000	25 000
Sud urbain	200	170	150	200	1,18	236	40 120	35 400
Sud rural	400	360	180	200	1,11	222	79 920	39 960
Total	1 000	730	500				160 040	109 110

Ainsi, la proportion estimative des ménages ayant accès aux soins de santé primaires est :

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109\ 110}{160\ 040} = 0{,}682, \text{ ou } 68{,}2\ \%$$

et le nombre estimatif de ménages ayant accès aux sois de santé primaires est :

$$\hat{t} = \sum w_h t_h = 109\ 110 = 68{,}2\ \% \text{ de } 160\ 040.$$

Il y a lieu de noter que la proportion estimative non pondérée de ménages ayant accès aux soins de santé primaires, sur la base uniquement des données fournies par les déclarants, est :

$$\hat{p}_{uw} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0{,}685, \text{ ou } 68{,}5\ \%$$

et la proportion estimative sur la base des pondérations initiales sans ajustement pour compenser la non-réponse est :

$$\hat{p}_1 = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83\ 000}{126\ 000} = 0{,}659, \text{ ou } 65{,}9\ \%.$$

40. Cet exemple a été fourni pour illustrer aussi comment les pondérations initiales sont ajustées pour compenser la non-réponse. Un écart notable existe entre la proportion estimée en utilisant uniquement les pondérations initiales et en utilisant les pondérations ajustées pour compenser la non-

réponse, mais la différence entre la proportion non pondérée et la proportion ajustée pour compenser la non-réponse semble négligeable.

41. Après ajustement des pondérations pour compenser la non-réponse, l'on peut, s'il y a lieu, apporter d'autres ajustements aux pondérations. L'on verra dans la section suivante comment sont ajustées les pondérations pour compenser la non-couverture.

6.6. Ajustement des pondérations d'échantillonnage pour compenser la non-couverture

42. Par non-couverture, l'on entend le fait que le cadre d'échantillonnage ne couvre pas l'intégralité de la population cible, de sorte que certaines unités de la population n'ont aucune probabilité d'être sélectionnées dans l'échantillon. Cela n'est qu'une des nombreuses déficiences possibles des cadres d'échantillonnage utilisés pour sélectionner des échantillons. Pour plus amples détails concernant les cadres d'échantillonnage, voir le chapitre 4.

43. La non-couverture suscite un problème majeur dans le contexte des enquêtes sur les ménages, surtout celles qui sont réalisées dans les pays en développement. L'impact de la non-couverture ressort clairement du fait que les estimations des effectifs de la population fondées sur certaines enquêtes réalisées dans les pays en développement sont bien inférieures à celles provenant d'autres sources. Par conséquent, les offices nationaux de statistique doivent, dans leurs programmes de travail et dans leurs activités de formation, concentrer leurs efforts sur les méthodes à utiliser pour identifier, évaluer et limiter la non-couverture dans le contexte des enquêtes sur les ménages.

44. La présente section évoque certaines des causes de la non-couverture dans les enquêtes sur les ménages et l'une des procédures utilisées pour la compenser, à savoir l'ajustement statistique des pondérations par une post-stratification.

6.6.1. Causes de la non-couverture dans les enquêtes sur les ménages

45. Dans les pays en développement, la plupart des enquêtes sur les ménages sont fondées sur une conception probabiliste stratifiée à plusieurs phases. Les unités sélectionnées lors de la première phase, c'est-à-dire les unités primaires d'échantillonnage, sont habituellement des unités géographiques. Lors de la deuxième phase, il est établi une liste des ménages ou des logements à partir de laquelle est sélectionné l'échantillon de ménages. Lors de la dernière phase, il est établi une liste de membres des ménages ou de résidents à partir de laquelle est sélectionné l'échantillon de personnes. Il peut donc y avoir non-couverture dans l'une quelconque de ces trois phases : au niveau des UPE, au niveau des ménages et au niveau des personnes.

46. Comme les UPE sont généralement sélectionnées sur la base des zones d'énumération identifiées et utilisées lors du précédent recensement de la population et du logement, elles sont censées couvrir tout le secteur géographique où se trouve la population cible. Ainsi, la non-couverture est généralement réduite dans le cas des UPE. Dans le contexte des enquêtes sur les ménages menées dans les pays en développement, la non-couverture au niveau des UPE ne suscite pas le même problème que la non-couverture aux étapes suivantes. Cependant, il n'en demeure pas moins une non-couverture d'UPE dans la plupart des enquêtes. Par exemple, même si une enquête est conçue de manière à fournir des estimations concernant l'ensemble de la population d'un pays ou d'une région du pays, certaines UPE peuvent être délibérément exclues au stade de la conception, par exemple si certaines régions sont inaccessibles par suite de troubles civils ou d'une catastrophe naturelle. En

outre, les régions reculées qui ne comportent qu'un très petit nombre de villages ou d'habitants sont parfois exclues des cadres d'échantillonnage car il serait trop coûteux d'y mener l'enquête; comme elles ne représentent qu'une faible proportion de la population, l'impact de leur exclusion sur les effectifs de la population est très réduit (voir le chapitre 4 pour plus amples informations et pour de nombreux exemples de non-couverture des UPE dans les enquêtes sur les ménages). L'exclusion de ces régions doit être expressément indiquée dans le rapport d'enquête. En effet, il ne faut pas donner l'impression que les résultats de l'enquête s'appliquent à l'ensemble du pays ou de la région alors qu'une partie de la population n'est pas couverte. Toutes les informations voulues concernant la non-couverture doivent être indiquées dans le rapport d'enquête.

47. La non-couverture devient un problème plus sérieux au niveau des ménages. La plupart des enquêtes considèrent les ménages comme un groupe de personnes qui sont habituellement apparentées d'une façon ou d'une autre et qui résident habituellement dans le même logement. Il reste d'importants problèmes de définition à résoudre, comme la question de savoir qui doit être considéré comme résident habituel et ce qu'il faut entendre par logement. Comment doivent être traitées des structures comportant plusieurs unités (comme un immeuble d'appartements) et des logements où vivent plusieurs ménages ? Il peut être facile d'identifier le logement mais difficile, si les structures sociales sont complexes, d'identifier quels sont les ménages qui vivent dans les logements en question. Il y a donc de sérieux risques d'erreur d'interprétation ou de manque de cohérence dans l'interprétation de ces concepts par des enquêteurs différents ou dans des cultures ou des pays différents. Quoi qu'il en soit, des instructions rigoureuses doivent être élaborées pour indiquer aux enquêteurs quelles sont les personnes qui doivent être considérées comme faisant partie d'un ménage et ce qui doit être considéré comme un logement.

48. Les autres facteurs qui contribuent à la non-couverture sont notamment l'omission involontaire de logements des listes établies lors des opérations sur le terrain ou de sous-groupes de population présentant un intérêt aux fins de l'enquête (par exemple des enfants en bas âge ou des personnes âgées), les omissions dues à des erreurs de mesure, l'exclusion de nombre de ménages se trouvant absents et les omissions imputables à une mauvaise compréhension des concepts à la base de l'enquête. Il y a également une dimension temporelle à ce problème : autrement dit, il se peut que les logements soient inoccupés ou en construction lors de l'établissement de la liste mais soient habités au moment de la collecte de données. Dans les pays en développement, le problème lié à la non-couverture est aggravé par le fait que, généralement, leurs recensements, qui sont la seule base disponible pour l'établissement de cadres d'échantillonnage, ne fournissent pas d'adresses détaillées des unités d'échantillonnage au niveau des ménages et des individus. Fréquemment, il est utilisé pour établir une liste des ménages des registres administratifs anciens ou inexacts et des membres du ménage se trouvent délibérément ou accidentellement exclus de la liste des résidents. L'on trouvera de plus amples détails sur les causes de non-couverture dans Lepkowski (2003) et dans les références citées dans cet ouvrage.

6.6.2. Compensation de la non-couverture dans les enquêtes sur les ménages

49. Plusieurs approches peuvent être envisagées pour faire face au problème lié à la non-couverture dans les enquêtes sur les ménages (Lepkowski, 2003), notamment celles qui consistent à :

a) Améliorer les procédures de terrain, comme l'utilisation de cadres multiples et de meilleures procédures d'établissement des listes;

b) Compensation de la non-couverture au moyen d'un ajustement statistique des pondérations.

50. S'agissant de la deuxième approche, si des totaux témoins fiables sont disponibles pour l'ensemble de la population ou pour les sous-groupes spécifiés de la population, l'on peut essayer d'ajuster les pondérations des unités d'échantillonnage de manière que la somme des pondérations corresponde aux totaux témoins des sous-groupes spécifiés. Ces sous-groupes et la procédure statistique d'ajustement sont appelés *post-stratification*. Cette procédure compense la non-couverture en ajustant la répartition pondérée d'échantillonnage pour certaines variables de manière qu'elle soit conforme à la répartition connue de la population [pour quelques exemples pratiques de la marche à suivre pour analyser les données post-stratification, voir Lehtonen et Pahkinen (1995)]. L'on en trouvera ci-après une illustration simple.

Exemple

Supposons que, dans l'exemple précédent, l'on sache, au moyen d'une source indépendante comme un registre de l'état civil à jour, qu'il y a 45 025 ménages dans le nord et 115 800 dans le sud. Supposons en outre que les totaux pondérés de l'échantillon soient respectivement de 40 000 et de 120 040. Il y a alors lieu de procéder en deux étapes, comme suit :

- *Étape 1.* Calcul des facteurs post-stratification :
  Pour la région nord, nous avons : $w_{3h} = \frac{45\,025}{40\,000} = 1{,}126$; et
  Pour la région sud, nous avons : $w_{3h} = \frac{115\,800}{120\,040} = 0{,}965$.
- *Étape 2.* Calcul de la pondération finale ajustée : $w_f = w_h \times w_{3h}$:
  Les résultats numériques sont résumés au tableau 6.4.

Tableau 6.4
Pondération post-stratifiée visant à compenser la non-couverture

Strate	r_h	t_h	w_h	w_{fh}	$w_{fh}\, r_h$	$w_{fh}\, t_h$
Nord urbain	80	70	125	140,75	11 260	9 852
Nord rural	120	100	250	281,40	33 768	28 140
Sud urbain	170	150	236	227,77	38 721	34 166
Sud rural	360	180	222	214,20	77 112	38 556
Total	730	500			160 861	110 714

La proportion estimative de ménages ayant accès aux soins de santé primaires est :

$$\hat{p}_f = \frac{\sum w_{fh} t_h}{\sum w_{fh}\, r_h} = \frac{110\,714}{160\,861} = 0{,}69, \text{ ou } 69\ \%.$$

51. Il y a lieu de noter que, lorsque les pondérations sont ajustées par post-stratification, les effectifs pondérés de l'échantillon pour les régions nord et sud sont respectivement de 45 024 (11 256 + 33 768) et 115 821 (38 709 + 77 112), chiffres qui sont très proches des totaux témoins indépendants susmentionnés.

6.7. Accroissement de la variance d'échantillonnage dû à la pondération

52. Alors même que l'utilisation de pondérations pour l'analyse des données provenant de l'enquête tend à réduire la distorsion des estimations, elle peut également accroître les variances de ces estimations. Pour simplifier, prenons le cas d'une conception stratifiée à une seule phase, les échantillons étant sélectionnés sur la base d'une probabilité égale à l'intérieur des différentes strates. Si les variances des strates (c'est-à-dire les variances entre unités des strates) ne sont pas les mêmes pour chaque strate, utiliser des pondérations inégales d'une strate à l'autre (par exemple des pondérations inversement proportionnelles aux variances des strates) peut donner des estimations plus précises. Cependant, si les variances des strates sont identiques pour chaque strate, des pondérations inégales entraîneront des variances plus marquées que s'il avait été utilisé des pondérations égales.

53. L'utilisation de pondérations a pour effet d'accroître la variance d'une moyenne estimative de la population par le facteur :

$$L = n \times \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2} \tag{6.12}$$

où $n = \sum_h n_h$

est la taille totale de l'échantillon réalisé, w_h est la pondération finale et n_h est la taille de l'échantillon réalisé pour la strate h. Cette formule peut également être présentée comme suit en termes de coefficient de variation des pondérations :

$$L = n \times \frac{\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j) \tag{6.13}$$

$$\text{où } CV^2(w_j) = \frac{n}{(\sum_j w_j)^2}\left\{\sum_j w_j^2 - \frac{1}{n}(\sum_j w_j)\right\} = \frac{\text{Variance des pondérations}}{(\text{moyenne des pondérations})^2}$$

Exemple

Nous allons maintenant calculer le facteur d'inflation de la variance en utilisant les données de l'exemple figurant dans la section 6.6.2, avec les pondérations finales w_{fh} et les tailles de l'échantillon réalisé r_h (voir le tableau 6.5).

Tableau 6.5
Paramètres de variance par strate

Strate	r_h	w_{fh}	$w_{fh} r_h$	$w_{fh}^2 r_h$
Nord urbain	80	140 75	11 260	1 584 845
Nord rural	120	281 40	33 768	9 502 315
Sud urbain	170	227 77	38 721	8 819 459
Sud rural	360	214 20	77 112	16 517 390
Total	**730**		**160 861**	**36 424 009**

$$\text{Donc, } L = 730 \times \frac{36\,424\,009}{(160\,861)^2} = 1{,}03.$$

Autrement dit, la variance des estimations issues de l'enquête augmente d'environ 3 % par suite de l'utilisation de pondérations.

6.8. Allégement des pondérations

54. Une fois que les pondérations ont été calculées et ajustées pour compenser les imperfections dont il est question plus haut, il est bon d'examiner la répartition des pondérations ajustées. Des pondérations extrêmement importantes, même si elles n'affectent qu'une faible proportion des cas sélectionnés, peuvent accroître considérablement la variance des estimations. La pratique commune consiste par conséquent à alléger les pondérations extrêmes pour les ramener à une valeur maximum prédéterminée afin de limiter la variation concomitante des pondérations (et ainsi de réduire la variance des estimations) tout en empêchant qu'un petit nombre d'unités sélectionnées ne domine l'estimation globale. Les pondérations sont le plus souvent allégées après l'ajustement visant à compenser la non-réponse.

55. L'allégement des pondérations, tout en tendant à réduire la variance des estimations, y introduit également une distorsion. Dans certains cas, la réduction de la variance due à l'allégement de très fortes pondérations peut plus que compenser l'aggravation de la distorsion introduite et réduire ainsi l'erreur carrée moyenne des estimations. Dans la pratique, les pondérations ne doivent être allégées que lorsqu'il y a des raisons de le faire, c'est-à-dire lorsqu'il peut être établi que la distorsion introduite par suite de l'utilisation de pondérations allégées (par opposition aux pondérations initiales) a moins d'impact sur l'erreur carrée moyenne totale que la réduction correspondante de la variance obtenue grâce à l'allégement.

56. Dans le cas d'une conception stratifiée, le processus d'allégement de pondération devrait idéalement être appliqué à l'intérieur de chaque strate. L'on commence par spécifier une unité supérieure pour les pondérations initiales avant d'ajuster toute la série des pondérations de sorte que la somme des pondérations allégées soit identique à celle des pondérations initiales. Disons que w_{hi} dénote la pondération finale de la $i^{ième}$ unité de la strate h et w_{hB} dénote la limite supérieure des pondérations spécifiques pour la strate h. Ainsi, la pondération allégée pour la $i^{ième}$ unité sélectionnée de la strate h peut être définie comme :

$$w_{hi(T)} = \begin{cases} w_{hi} & \text{si } w_{hi} < w_{hB} \\ w_{hB} & \text{si } w_{hi} \geq w_{hB} \end{cases} \qquad (6.14)$$

57. Les pondérations allégées pour l'ensemble de l'échantillon peuvent maintenant être ajustées de nouveau de sorte que leur somme soit exactement identique à celle des pondérations initiales. Dans un souci de simplicité, nous supposerons que les pondérations sont constantes à l'intérieur de chaque strate et abandonnerons la liste i pour le reste de la discussion. Disons que F_T dénote le ratio entre la somme des pondérations initiales et la somme des pondérations allégées, autrement dit :

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}} \tag{6.15}$$

où les sommes du ratio sont reportées sur toutes les strates et par conséquent sur toutes les unités d'échantillonnage. Si nous définissions la pondération allégée ajustée pour la $h^{ième}$ strate comme étant :

$$w^*_{h(T)} = F_{Th} \times w_{(T)} \tag{6.16}$$

il est clair que $\sum_h n_h w^*_{h(T)} = \sum_h n_h w_h$, comme souhaité.

L'exemple suivant illustrera et fera comprendre plus clairement la procédure d'allégement.

58. Les deux premières colonnes du tableau 6.6 ci-après indiquent le nombre total d'unités et la pondération finale, respectivement, pour chacune des sept strates. Il a été choisi une pondération maximum de 250 de sorte que les pondérations initiales sont tronquées à 250, comme le montre la troisième colonne du tableau.

Tableau 6.6
Allégement des pondérations

	n_h	w_h	$w_{h(T)}$	$n_h w_h$	$n_h w_{h(T)}$	$n_h w^*_{h(T)}$
	80	140,75	140,75	11 260	11 260	11 823,00
	100	150,25	150,25	15 025	15 025	15 776,25
	125	175,00	175,00	21 875	21 875	22 968,75
	150	200,00	200,00	30 000	30 000	31 500,00
	120	250,00	250,00	30 000	30 000	31 500,00
	120	275,13	250,00	33 015	30 000	31 500,00
	170	285,40	250,00	48 518	42 500	44 625,00
Total	**865**			**189 693**	**180 660**	**189 693,00**

Il y a lieu de noter que, dans ce cas,

$$F_T = \frac{\sum_i n_h w_{hi}}{\sum_i n_h w_{hi(T)}} = \frac{189\ 693}{180\ 660} = 1{,}05.$$

Les pondérations allégées ont été recalculées de sorte que leur somme soit équivalente au total initial 125, en multipliant chaque pondération par F_T =1,05.

6.9. Conclusion

59. La pondération d'échantillonnage est aujourd'hui considérée comme faisant partie intégrante de l'analyse des données provenant des enquêtes sur les ménages réalisées dans les pays en développement comme dans le reste du monde. La plupart des programmes d'enquête préconisent actuellement l'utilisation de pondérations même dans de rares cas d'échantillons autopondérés (en l'occurrence, les pondérations seraient égales à l'unité). Par le passé, les concepteurs ont fait d'énormes efforts pour atteindre l'objectif virtuellement irréalisable que sont des échantillons autopondérés afin d'éliminer ainsi la nécessité des pondérations lors de l'analyse des données provenant de l'enquête. L'idée communément reçue était que l'utilisation de pondérations compliquait à l'excès les analyses et que l'infrastructure nécessaire à une analyse pondérée était très réduite, voire inexistante. Cependant, les progrès de l'informatique réalisés au cours des dix dernières années ont ôté tout son poids à cet argument. Des ordinateurs et des logiciels sont aujourd'hui disponibles à peu de frais dans nombre de pays en développement. En outre, il a été mis au point beaucoup de logiciels spécialisés pour l'analyse des données provenant des enquêtes. Ces systèmes sont analysés et comparés au chapitre 7.

60. Comme on l'a dit, l'utilisation de pondérations réduit les distorsions dues aux imperfections de l'échantillon liées à la non-couverture et à la non-réponse. Les erreurs de non-réponse et de non-couverture sont des erreurs de types différents dues au fait que l'enquête n'a pas été conçue de manière à obtenir des informations de certaines unités de la population cible. Pour les enquêtes sur les ménages réalisées dans les pays en développement, la non-couverture est un problème plus sérieux que la non-réponse. L'on a donné dans ce chapitre des exemples des procédures à suivre pour identifier et ajuster statistiquement les pondérations de base de manière à compenser certains de ces problèmes et l'on a indiqué comment utiliser les pondérations ajustées pour estimer les paramètres à calculer. Aujourd'hui qu'il existe des ordinateurs rapides et des logiciels statistiques abordables ou gratuits, l'utilisation de pondérations devrait devenir systématique pour l'analyse des données provenant d'enquêtes sur les ménages, même dans les pays en développement. Toutefois, comme démontré dans ce chapitre, l'élaboration de pondérations d'échantillonnage complique à divers égards le déroulement des enquêtes. Par exemple, des pondérations doivent être calculées pour chacune des étapes de la sélection de l'échantillon; elles doivent ensuite être ajustées pour compenser les différentes imperfections de l'échantillon; et, enfin, elles doivent être conservées et utilisées à bon escient lors de toutes les analyses ultérieures. Il importe par conséquent d'accorder toute l'attention voulue aux opérations de pondérations et au calcul effectif des pondérations à utiliser pour l'analyse.

Références et autres lectures

Brick, J. M. et G. Kalton (1996). Handling missing data in survey research. *Statistical Methods in Medical Research,* vol. 5, p. 215-238.

Cochran, W. G. (1977). *Sampling Techniques,* troisième édition. New York, John Wiley & Sons.

Groves, R. M. et M. P. Couper (1998). *Non-response in Household Interview Surveys.* New York, John Wiley & Sons.

Groves, R. M., *et al.* (2002). *Survey Non-response.* New York, John Wiley & Sons.

D. Kasprzyk (1986). The treatment of missing survey data. *Survey Methodology,* vol. 12, p. 1-16.

Kalton, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor, Michigan, Survey Research Center, University of Michigan.

Kish, L. (1965). *Survey Sampling.* New York, Wiley.

I. Hess (1950). On non-coverage of sample dwellings, *Journal of the American Statistical Association,* vol. 53, p. 509-524.

Lehtonen, R. et E. J. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys.* New York, Wiley.

Lepkowski, James (2005). *L'erreur de non-observation dans les enquêtes sur les ménages dans les pays en développement.* Dans Enquêtes sur les ménages dans les pays en développement et les pays en transition. Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F. 05.XVII.6.

Lessler, J. et W. Kalsbeek (1992). *Nonsampling Error in Surveys.* New York, John Wiley & Sons.

Levy, P. S. et S. Lemeshow (1999). *Sampling of Populations: Methods and Applications,* troisième édition. New York, John Wiley & Sons.

Lohr, S. (1999). *Sampling: Design and Analysis,* Duxbury Press, Pacific Grove.

Yansaneh, I. S. (2004). *Aperçu des problèmes de conception d'échantillon pour les enquêtes sur les ménages dans les pays en développement et les pays en transition.* Dans Enquêtes sur les ménages dans les pays en développement et les pays en transition. Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F. 05.XVII.6.

Chapitre 7

Estimation des erreurs d'échantillonnage dans les données d'enquête

7.1. Introduction

1. Le présent chapitre donne un bref aperçu des diverses méthodes employées pour estimer les erreurs d'échantillonnage dans les données d'enquête sur les ménages provenant de différentes conceptions de l'échantillon, allant des conceptions standard que l'on trouve dans n'importe quel manuel élémentaire de la théorie de l'échantillonnage [par exemple, Cochran (1977)] à des conceptions plus complexes utilisées pour les enquêtes sur les ménages à grande échelle. Il offre pour les conceptions standard des formules et des exemples numériques illustrant l'estimation des erreurs d'échantillonnage, la construction des intervalles de confiance et les calculs des effets de conception et de la taille effective des échantillons. Il présente ensuite des méthodes d'estimation des erreurs d'échantillonnage pour des conceptions plus complexes. Les avantages et les inconvénients de chaque méthode sont examinés, et il est donné des exemples numériques pour illustrer son application. Il est fourni un exemple, basé sur des données provenant d'une enquête réelle, pour illustrer comment les logiciels statistiques standard sous-estiment les erreurs d'échantillonnage des estimations et conduisent ainsi à tirer des conclusions erronées concernant les paramètres étudiés. Pour éviter ce problème, il est vivement recommandé d'employer des logiciels statistiques spéciaux qui tiennent pleinement compte de la complexité des conceptions communément utilisées pour la réalisation d'enquêtes sur les ménages. Plusieurs de ces logiciels sont décrits et comparés.

7.1.1. Estimation des erreurs d'échantillonnage des données d'enquêtes complexes

2. Ces derniers temps, les objectifs d'enquêtes sur les ménages bien conçues ne consistent plus seulement à analyser les tableaux récapitulatifs des dénombrements ou des totaux concernant les paramètres étudiés. À l'heure actuelle, les analystes s'intéressent aussi à l'élaboration et à la mise à l'épreuve d'hypothèses ou à la modélisation. Par exemple, plutôt que de se borner à estimer la proportion d'une population qui vit au-dessous du seuil de pauvreté ou qui a un niveau d'instruction égal ou supérieur au secondaire, les analystes veulent maintenant évaluer l'impact des politiques ou étudier comment une variable d'intervention clé, par exemple les résultats scolaires d'un enfant, ou le niveau de pauvreté d'un ménage, est affectée par des facteurs comme la région, la situation socioéconomique, le sexe et l'âge.

3. Pour répondre à ces types de questions, il faut analyser en détail les données recueillies au niveau des ménages ou des individus. Inévitablement, la publication des résultats de telles analyses doit comprendre une indication des mesures permettant d'évaluer la précision et l'exactitude des estima-

tions tirées des données d'enquête. De telles informations sont indispensables si l'on veut pouvoir utiliser et interpréter comme il convient les résultats ainsi qu'évaluer et améliorer les conceptions et procédures d'échantillonnage. Ce suivi et cette évaluation des conceptions des échantillons sont particulièrement importants dans le cas de vastes programmes nationaux d'enquête, qui sont fréquemment conçus comme devant constituer la seule source d'informations détaillées sur des thèmes extrêmement divers.

4. L'une des principales mesures de la précision est la variance d'échantillonnage (dont le concept est présenté au chapitre 3), indicateur de la variabilité introduite par la décision de procéder par sondage plutôt que de recenser l'ensemble de la population, l'hypothèse étant que les informations rassemblées au moyen de l'enquête sont correctes. La variance d'échantillonnage est une mesure de la variabilité de la répartition d'une estimation. L'erreur type, qui est la racine carrée de la variance, est utilisée pour mesurer l'erreur d'échantillonnage. Quelle que soit l'enquête, l'estimation de cette erreur d'échantillonnage peut être évaluée et utilisée pour porter une appréciation sur l'exactitude des données.

5. La forme que revêt l'estimation de la variance et son évaluation dépendent de la conception de l'échantillon. Dans le cas des conceptions standard, ces estimations sont fréquemment évaluées au moyen de formules simples. Dans le cas des conceptions complexes utilisées pour les enquêtes sur les ménages, qui comportent souvent une stratification, une mise en grappes et des probabilités inégales de sélection, les formes de ces estimations sont fréquemment complexes et difficiles à évaluer. En pareil cas, il faut, pour calculer les erreurs d'échantillonnage, suivre des procédures qui tiennent compte de la complexité de la conception de l'échantillon utilisé pour générer les données, ce qui exige souvent, à son tour, le recours à des logiciels appropriés.

6. Dans beaucoup de pays en développement, l'analyse des données provenant d'enquêtes sur les ménages est souvent limitée à une analyse tabulaire de base, avec des estimations des moyennes, des proportions et des totaux, sans toutefois d'indications quant à la précision ou à l'exactitude de ces estimations. Même les offices nationaux de statistique disposant d'une solide infrastructure de collecte et de traitement de données statistiques manquent fréquemment d'expérience de l'analyse détaillée des données au niveau micro. Le concepteur ou l'analyste sera fréquemment surpris d'apprendre, par exemple, que la mise en grappes des éléments introduit entre ces derniers des corrélations qui réduisent la précision des estimations par rapport aux échantillons aléatoires simples qu'ils ont coutume d'analyser, ou que l'utilisation de pondérations dans l'analyse gonfle généralement les erreurs d'échantillonnage, ou encore que les logiciels standard qu'ils utilisent généralement dans leur travail ne tiennent pas compte comme il convient de cette moindre précision.

7. Le présent chapitre essaie de remédier à cette situation en donnant un bref aperçu des méthodes de calcul des erreurs d'échantillonnage pour les types de conceptions complexes habituellement employées pour les enquêtes sur les ménages dans les pays en développement ainsi que les logiciels statistiques utilisés pour l'analyse des données. Il est présenté plusieurs exemples numériques pour illustrer les procédures de variance évoquées.

7.1.2. Aperçu général

8. La section 7.2 donne une définition de la variance d'échantillonnage dans le cas d'un sondage aléatoire simple, de même que des exemples numériques illustrant le calcul de la variance d'échantillonnage et la construction des intervalles de confiance. La section 7.3 contient les définitions

d'autres mesures de l'erreur d'échantillonnage, et la section 7.4 des formules permettant de calculer la variance d'échantillonnage selon différentes conceptions standard, comme l'échantillonnage stratifié et l'échantillonnage en grappes. Il est fourni plusieurs exemples génériques pour faire mieux comprendre les concepts. La section 7.5 traite des caractéristiques usuelles des différentes conceptions des enquêtes sur les ménages ainsi que du contenu et de la structure des données qui sont nécessaires pour pouvoir estimer comme il convient l'erreur d'échantillonnage. La forme que revêtent généralement les estimations que cherchent à obtenir les enquêtes sur les ménages est également indiquée. La section 7.6 donne de brèves indications touchant la présentation des informations concernant les erreurs d'échantillonnage et la section 7.7 décrit les méthodes à utiliser dans la pratique pour calculer les erreurs d'échantillonnage dans le cas de conceptions très complexes. Ces méthodes exigent fréquemment le recours à des procédures spéciales et à des logiciels spécialisés. Les problèmes que peut soulever l'utilisation de logiciels statistiques standard pour l'analyse des donnés d'enquête sont discutés dans la section 7.8, qui contient également un exemple fondé sur les données recueillies lors d'une enquête sur la couverture des programmes de vaccination réalisés au Burundi en 1989. Certains logiciels disponibles dans le commerce qui permettent d'atténuer les erreurs d'échantillonnage sont examinés et comparés dans les sections 7.9 et 7.10. Ce chapitre s'achève sur quelques dernières observations, qui figurent dans la section 7.11.

7.2. Variance d'échantillonnage dans le cas d'un sondage aléatoire simple

9. La variance d'échantillonnage d'une estimation peut être définie comme étant le carré de l'écart moyen par rapport à la valeur moyenne des estimations, la moyenne étant calculée sur la base de tous les échantillons possibles. Comme indiqué au chapitre 3, l'échantillonnage aléatoire simple est la plus élémentaire des techniques d'échantillonnage, mais elle est rarement utilisée pour des enquêtes de grande envergure car son application est très peu efficace et d'un coût prohibitif.

10. Pour mieux comprendre le concept de variance d'échantillonnage, nous prendrons le cas d'une population restreinte de cinq ménages ($N = 5$), parmi laquelle sera sélectionné un petit échantillon de ménages ($n = 2$) par échantillonnage aléatoire simple sans remplacement. L'on supposera en outre que la variable visée est les dépenses mensuelles d'alimentation du ménage et que les dépenses de chacun des quatre ménages sont comme indiqué au tableau 7.1 ci-après :

Tableau 7.1
Dépenses mensuelles d'alimentation par ménage, en dollars

Ménage	Dépenses d'alimentation, en dollars (Y_1)
1	10
2	20
3	30
4	40
5	50

11. Il y a lieu de noter d'abord qu'étant donné que nous connaissons la valeur de la variable étudiée pour tous les ménages faisant partie de notre population nous pouvons calculer la valeur des paramètres correspondant aux dépenses mensuelles moyennes d'alimentation par ménage, c'est-à-dire :

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{10+20+30+40+50}{5} = \frac{150}{5} = 30.$$

L'estimation de l'échantillonnage aléatoire simple sans remplacement pour les dépenses mensuelles d'alimentation est :

$$\hat{\bar{Y}} = \frac{1}{2}\sum_{i \varepsilon S} Y_i,$$

la somme représentant les unités sélectionnées. Manifestement, l'estimation obtenue dépend de l'échantillon sélectionné. Le tableau 7.2 ci-dessous illustre tous les échantillons possibles, l'estimation résultant de chaque échantillon, les écarts de chaque estimation par rapport à la moyenne de la population et les écarts au carré. Il y a lieu de noter que $\hat{\bar{Y}}_{ave}$ dénote la moyenne de toutes les estimations fondées sur les échantillons et que $\bar{Y}$ est le symbole de la moyenne de la population et que $\hat{\bar{Y}}$ est le symbole de l'estimation de la moyenne de la population, appelée moyenne de l'échantillon.

Tableau 7.2
Calcul de la variance d'échantillonnage réelle de $\hat{\bar{Y}}$, paramètre pour la moyenne

Échantillon	Unités d'échantillonnage	Estimation de l'échantillon ($\hat{\bar{Y}}_i$)	$\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave}$	$(\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave})^2$
1	(1, 2)	15	–15	225
2	(1, 3)	20	–10	100
3	(1, 4)	25	–5	25
4	(1, 5)	30	0	0
5	(2, 3)	25	–5	25
6	(2, 4)	30	0	0
7	(2, 5)	35	5	25
8	(3, 4)	35	5	25
9	(3, 5)	40	10	100
10	(4, 5)	45	15	225
Moyenne		**30**	**0**	**750**

Il y a lieu de noter que la moyenne des estimations fondées sur tous les échantillons possibles est :

$$\hat{\bar{Y}}_{ave} = \frac{1}{10}\sum_{i=1}^{10} \hat{\bar{Y}}_i = \frac{15+20+25+30+25+30+35+35+40+45}{10} = \frac{300}{10} = 30 = \bar{Y}.$$

12. Autrement dit, la valeur moyenne des estimations pour tous les échantillons possibles est égale à la moyenne de la population. Une estimation présentant une telle caractéristique est appelée estimation *sans distorsion* du paramètre à l'étude.

13. La valeur d'échantillonnage réelle de l'estimation des dépenses mensuelles moyennes d'alimentation sur la base d'un échantillon aléatoire simple sans remplacement de taille $n = 2$ pour cette population est :

$$Var(\hat{\bar{Y}}) = \frac{1}{10}\sum_{i=1}^{10} (\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave})^2 = \frac{750}{10} = 75.$$

14. Le problème que soulève cette approche tient au fait qu'il n'est pas possible de sélectionner tous les échantillons possibles de la population dont il s'agit. Dans la pratique, il est sélectionné un seul échantillon, et les valeurs propres à la population visée ne sont pas connues. Une méthode plus réaliste consiste à utiliser des formules pour calculer la variance. Il existe de telles formules pour toutes les conceptions standard.

15. Dans le cas d'un échantillon aléatoire simple sans remplacement, la variance d'échantillonnage d'une moyenne estimative ($\hat{\bar{Y}}$), sur la base d'un échantillon de taille *n,* est donnée par l'expression :

$$Var(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{\delta^2}{n}. \qquad (7.1)$$

où $\delta^2 = \dfrac{\sum_{i=1}^{N}(Y_i - \hat{\bar{Y}})^2}{N-1}$

est une mesure de la variabilité de la caractéristique à l'étude (variance de population de *Y*). Habituellement, δ^2 est inconnue et doit être estimée à partir de l'échantillon (voir l'équation 7.2 ci-dessous). Il ressort clairement de la formule (7.1) que la variance d'échantillonnage dépend des facteurs ci-après :

a) Variance de la population et de la caractéristique à l'étude;
b) Effectifs de la population;
c) Taille de l'échantillon;
d) Conception de l'échantillon et méthode de l'estimation.

16. La proportion de la population faisant partie de l'échantillon, *n/N,* est appelée la fraction d'échantillonnage (dénotée par *f*) et le facteur [1 – (*n/N)*], ou *1 – f,* qui est la proportion de la population ne faisant pas partie de l'échantillonnage, est appelé coefficient de correction de la population finie (*fpc*). Le *fpc* représente l'ajustement apporté à l'erreur type de l'estimation pour tenir compte du fait que l'échantillon est sélectionné sans remplacement parmi une population finie. Il y a lieu de noter toutefois que, lorsque la fraction d'échantillonnage est finie, le *fpc* peut être ignoré. Dans la pratique, le *fpc* peut être ignoré s'il ne dépasse pas 5 % (Cochran, 1977).

17. La formule ci-dessus montre que la variance d'échantillonnage est inversement proportionnelle à la taille de l'échantillon. À mesure que la taille de l'échantillon augmente, la variance d'échantillonnage diminue et, dans le cas d'un recensement ou d'une énumération complète (où $n = N$), il n'y a pas de variance d'échantillonnage. Il convient de noter que la non-réponse a pour effet, dans la pratique, de réduire la taille de l'échantillon de sorte qu'elle accroît la variabilité de l'échantillonnage.

18. Une estimation dépourvue de distorsion de la valeur d'échantillonnage de la moyenne estimative est donnée par :

$$v(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{s^2}{n} \qquad (7.2)$$

où s^2 est une estimation de la variance de population, δ^2, sur la base de l'échantillon. C'est ce que l'on appelle la variance de l'échantillon. L'intervalle de confiance de 95 % pour la moyenne de la population (voir le paragraphe 30 du chapitre 3) est donné par la formule :

$$\hat{\bar{Y}} \pm 1{,}96\sqrt{v(\hat{\bar{Y}})} \tag{7.3}$$

19. Pour une proportion seulement de la population, l'estimation de l'échantillon et la variance estimative sont données respectivement par :

$$\hat{P} = \frac{\textit{nombre d'unités présentant la caractéristique visée}}{n} \tag{7.4}$$

$$\text{et } v(\hat{P}) = (1 - \frac{n}{N})\frac{\hat{P}(1-\hat{P})}{n-1} \tag{7.5}$$

20. Le tableau 7.3 ci-dessous résume les estimations de différentes quantités de la population et les variances des estimations dans le cas d'un échantillon aléatoire simple sans remplacement.

Tableau 7.3
Estimations et leurs variances pour les caractéristiques de population sélectionnées

Paramètre	Estimation	Variance de l'estimation
Moyenne de la population ($\hat{\bar{Y}}$)	$\hat{\bar{Y}} = \frac{1}{n_{\text{éc}}} \sum_{\text{Échantillon}} Y_i$	$v(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{s^2}{n}$
Total de la population	$\hat{T} = N\hat{\bar{Y}}$	$v(\hat{T}) = N^2 v(\hat{\bar{Y}})$
Proportion de la population pour une catégorie	$\hat{P} = \frac{\textit{Nombre d'unités sélectionnées dans la catégorie}}{n}$	$v(\hat{P}) = (1 - \frac{n}{N})\frac{\hat{P}(1-\hat{P})}{n-1}$

21. D'une manière générale, l'intervalle de confiance de $(1-\alpha)$ pour la moyenne de la population est donné par :

$$\textit{Estimation} \pm z_{1-\alpha/2}\sqrt{\textit{variance estimative de l'estimation}} \tag{7.6}$$

où $z_{1-\alpha/2}$ est le $(1-\alpha/2)$ $n^{ième}$ percentile de la répartition normale type.

22. L'exemple ci-après illustre l'estimation de la variance d'échantillonnage sur la base d'un échantillon sélectionné.

Exemple 1

Prenons un échantillon aléatoire simple de n = 20 ménages sélectionnés parmi une population nombreuse de N = 20 000 ménages. Les données rassemblées sont présentées dans le tableau 7.4 ci-après, où la variable Y dénote les dépenses hebdomadaires d'alimentation des ménages et la variable Z la possession d'un poste de télévision (z = 1 si oui, sinon 0).

Tableau 7.4
Dépenses hebdomadaires d'alimentation des ménages et possession d'un poste de télévision parmi les ménages sélectionnés

Ménage	Y_i	Z_i	i	Y_i	Z_i
1	5	0	11	7	1
2	10	1	12	8	1
3	5	0	13	9	1
4	9	1	14	10	1
5	5	1	15	8	1
6	6	1	16	8	0
7	7	0	17	5	0
8	15	1	18	7	0
9	12	1	19	12	1
10	8	0	20	4	0
Total		160	12		

L'estimation des dépenses mensuelles d'alimentation pour la moyenne de la population est :

$$\hat{\bar{Y}} = \frac{1}{20}\sum_{i=1}^{20} Y_i = \frac{5+10+\ldots+12+4}{20} = \frac{160}{20} = 8.$$

L'estimation de la variance de la moyenne estimative est :

$$v(\hat{\bar{Y}}) = (1-\frac{20}{20\ 000})\left\{\frac{(5-8)^2+(10-8)^2+\ldots..+(12-8)^2+(4-8)^2}{19}\right\} = 7{,}87.$$

L'intervalle de confiance de 95 % pour la moyenne de la population est :

$$8 \pm 1{,}96 \times \sqrt{7{,}87} = (2{,}50,\ 13{,}50).$$

L'estimation des dépenses mensuelles d'alimentation des ménages pour le total de la population est :

$$\hat{Y} = N\hat{\bar{Y}} = 20\ 000 \times 8 = 160\ 000.$$

L'estimation de la variance du total estimatif est :

$$v(\hat{Y}) = 20\ 000^2 \times 8{,}87 = 3\ 148\ 000\ 000.$$

L'intervalle de confiance de 95 % pour les dépenses de la population est :

$$160\ 000 \pm 1{,}96 \times \sqrt{3\ 148\ 000\ 000} = (50\ 030,\ 269\ 970).$$

L'estimation de la proportion de la population de ménages qui possèdent un poste de télévision est : $\hat{P} = \frac{1}{20}\sum_{i=1}^{20} Z_i = \frac{12}{20} = 0{,}6.$

L'estimation de la variance de la population estimative des ménages ayant un poste de télévision est :

$$v(\hat{P}) = (1-\frac{20}{20\ 000})\frac{0{,}6(1-0{,}6)}{19} = 0{,}0126.$$

L'intervalle de confiance de 95 % pour la moyenne de la population est :

$0{,}6 \pm 1{,}96 \times \sqrt{0{,}0126} = (0{,}38,\ 0{,}82)$.

7.3. Autres mesures de l'erreur d'échantillonnage

23. Indépendamment de la variance d'échantillonnage, il y a d'autres mesures de l'erreur d'échantillonnage, dont l'erreur type, le coefficient de variation et l'effet de conception. Ces mesures sont algébriquement liées en ce sens que l'on peut dériver l'expression de l'une quelconque d'entre elles à partir des autres au moyen de simples opérations algébriques.

7.3.1. Erreur type

24. L'erreur type d'une estimation est la racine carrée de sa variance d'échantillonnage. Cette mesure est plus facile à interpréter étant donné qu'elle donne une indication de l'erreur d'échantillonnage sur la même échelle que l'estimation, tandis que la variance est fondée sur le carré des différences.

25. Une question qui se pose fréquemment lors de la conception d'une enquête est de savoir quelle est l'erreur type maximum qui peut être considérée comme tolérable. La réponse à cette question dépend de l'ampleur de l'estimation. Par exemple, une erreur type égale à 100 serait considérée comme modeste dans le contexte d'une estimation du revenu annuel mais importante lors d'une estimation du poids moyen des individus. Ainsi, dans l'exemple 1 ci-dessus, l'erreur type de $\sqrt{3\ 148\ 000\ 000} = 56\ 107$ pour un total estimatif de 160 000 peut être considérée comme trop importante.

7.3.2. Coefficient de variation

26. Le coefficient de variation (CV) d'une estimation est le ratio entre son erreur type et la valeur moyenne de l'estimation elle-même. Ainsi, le CV constitue une mesure de l'erreur d'échantillonnage par rapport à la caractéristique étudiée. Il est habituellement exprimé sous forme de pourcentage.

27. Le CV est utile pour comparer la précision de l'estimation de tailles ou d'échelles différentes, mais il n'est guère utile dans le cas d'estimations de caractéristiques dont la valeur réelle peut être nulle ou négative, par exemple des estimations d'un changement, comme l'évolution du revenu moyen sur une période de deux ans.

7.3.3. Effet de conception

28. L'effet de conception (appelé *deff*) est défini comme étant le ratio entre la variance d'échantillonnage d'une estimation dans le cas d'une conception donnée et la variance d'échantillonnage de l'estimation fondée sur un échantillon aléatoire simple de même taille. Il s'agit en quelque sorte du coefficient par lequel la variance d'une estimation fondée sur un échantillon aléatoire simple de même taille doit être multipliée pour tenir compte des complexités de la conception effective de l'échantillon dues à des facteurs comme la stratification, la mise en grappes et la pondération.

29. Autrement dit, une estimation fondée sur des données provenant d'un échantillon complexe de taille n a la même variance que l'estimation calculée à partir de données provenant d'un échantillon aléatoire simple de taille $n/deff$. Par conséquent, le ratio $n/deff$ est parfois appelé taille effective de

l'échantillon dans le cas d'estimations fondées sur des données provenant d'une conception complexe. Pour une discussion générale du calcul de la « taille effective de l'échantillon », voir Kish (1995) et Potthoff, Woodbury et Manton (1992) et les références citées par ces auteurs. Voir également les différentes sections du chapitre 3 pour une analyse plus détaillée des efforts de conception et de leur utilisation dans la conception de l'échantillon.

7.4. Calcul de la variance d'échantillonnage pour d'autres conceptions standard

30. Dans le cas de conceptions simples et d'estimations linéaires simples comme des moyennes, des proportions et des totaux, l'on peut habituellement dériver des formules pouvant servir à calculer les variances des estimations. Cependant, pour les types de conceptions et d'estimations complexes qui caractérisent habituellement les enquêtes sur les ménages, cela est souvent difficile, voire impossible. L'on trouvera dans la présente section des exemples illustrant le calcul de la variante d'échantillonnage pour un échantillonnage stratifié et un échantillonnage en grappes à une seule étape. Les manuels (par exemple Cochran, 1977, et Kish, 1965) contiennent des formules et des exemples de calcul des variances pour d'autres conceptions d'échantillonnage standard.

7.4.1. Échantillonnage stratifié

31. Le chapitre 3 contient une description détaillée de l'échantillonnage stratifié. Dans la présente section, nous ferons seulement porter notre attention sur l'estimation de la variance. Prenons le cas d'une conception stratifiée comportant H strates, les estimations d'échantillonnage des moyennes de la population des strates étant données par $\bar{Y}_1, \bar{Y}_2, \ldots \ldots \bar{Y}_H$, et les estimations d'échantillonnage des variances de population pour les strates par $S_1^2, S_2^2, \ldots \ldots S_H^2$. Selon cette conception, une estimation de la moyenne de la population est :

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{H} \hat{\bar{Y}}_h \qquad (7.7)$$

où $\hat{\bar{Y}}_h$ est l'estimation fondée sur l'échantillon de $\bar{Y}_h$, $h = 1, \ldots \ldots H$. La variance de l'estimation est donnée par :

$$v(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{H} v(\hat{\bar{Y}}_h) \qquad (7.8)$$

Dans le cas d'un échantillonnage aléatoire stratifié, l'estimation et sa variance estimative sont données par :

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} W_h \bar{y}_h \qquad (7.9)$$

où $\bar{y}_h$ est la moyenne de l'échantillon pour la strate h, N_h la taille de la population de la strate h, et $W_h = \frac{N_h}{N}$, $h = 1, \ldots \ldots H$.

La variance estimative de cette estimation dans le cas d'un échantillonnage aléatoire stratifié est donnée par :

$$v(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{H} W_h^2 v(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1 - \frac{n_h}{N_h}) \frac{s_h^2}{n_h} \quad (7.10)$$

où n_h est la taille de l'échantillon de la strate h et s_h^2 la variance de l'échantillon, estimation basée sur l'échantillon de S_h^2, $h = 1, \ldots \ldots H$.

Exemple 2

Nous allons maintenant appliquer ces résultats à un exemple de conception stratifiée comportant trois strates caractérisées par des paramètres tels que ceux qui sont indiqués dans le tableau 7.5 ci-dessous. Supposons que nous souhaitions estimer la moyenne de la population sur la base d'un échantillon global de 1 500 unités.

Tableau 7.5
Exemple de données pour une conception d'échantillon stratifié

Paramètre	Population	Strate 1 (capitale)	Strate 2 (province-milieu urbain)	Strate 3 (province-milieu urbain)
Taille	$N = 1\,000\,000$	$N_1 = 300\,000$	$N_2 = 500\,000$	$N_3 = 200\,000$
Variance	$S^2 = 75\,000$	$S_1^2 = ?$	$S_2^2 = ?$	$S_3^2 = ?$
Moyenne	$\bar{Y} = ?$	$\bar{Y}_1 = ?$	$\bar{Y}_2 = ?$	$\bar{Y}_3 = ?$
Coût unitaire	N/A	$C_1 = 1$	$C_2 = 4$	$C_3 = 16$
Taille de l'échantillon selon l'allocation optimale[a]	$n = 1\,500$	$n_1 = 857$	$n_2 = 595$	$n_3 = 48$
Moyenne de l'échantillon	N/A	$\bar{y}_1 = 4\,000$	$\bar{y}_2 = 2\,500$	$\bar{y}_3 = 1\,000$
Variance de l'échantillon	N/A	$s_1^2 = 90\,000$	$s_2^2 = 62\,500$	$s_3^2 = 10\,000$

Note : N/A signifie non applicable.
[a] Voir le chapitre 3.

L'estimation de la moyenne de la population est :

$$\hat{\bar{Y}}_{st} = \frac{300\,000}{1\,000\,000} \times 4\,000 + \frac{500\,000}{1\,000\,000} \times 2\,500 + \frac{200\,000}{1\,000\,000} \times 1\,000 = 2\,650.$$

La variance estimative de l'estimation ci-dessus est :

$$v(\hat{\bar{Y}}_{st}) = \left(\frac{300\,000}{1\,000\,000}\right)^2 \left(1 - \frac{857}{300\,000}\right)\left(\frac{90\,000}{857}\right) + \left(\frac{500\,000}{1\,000\,000}\right)^2 \left(1 - \frac{595}{500\,000}\right)\left(\frac{62\,500}{595}\right) +$$
$$\left(\frac{200\,000}{1\,000\,000}\right)^2 \left(1 - \frac{48}{200\,000}\right)\left(\frac{10\,000}{48}\right) = 43{,}98516.$$

L'intervalle de confiance de 95 % pour la moyenne de la population est :

$$2\,650 \pm 1\,96 \times \sqrt{43{,}98516} = (2\,637,\ 2\,663).$$

Il y a lieu de noter que la variance estimative de la moyenne estimative dans le cas d'un échantillonnage aléatoire simple est donnée par :

$$v(\hat{\bar{Y}}_{SRS}) = \left(1 - \frac{1\,500}{1\,000\,000}\right) \times \frac{75\,000}{1\,500} = 49{,}925.$$

Par conséquent, l'effet de conception de cette conception stratifiée est de $\frac{43{,}98516}{49{,}925} = 0{,}88$ et la taille effective de l'échantillon est de $\frac{1\,500}{0{,}88} = 1\,705$.

Cela signifie que l'estimation fondée sur un échantillon aléatoire stratifié de 1 500 unités a la même variance que celle qui est fondée sur un échantillon aléatoire simple de 1 705 unités.

32. Le chapitre 3 contient une description détaillée de la méthode d'échantillonnage en grappes. Dans la présente section, nous donnerons un seul exemple pour illustrer le calcul des erreurs d'échantillonnage dans le cas particulier d'un échantillonnage en grappes à une seule étape.

Exemple 3

Supposons que nous souhaitions estimer la proportion d'enfants en âge de fréquenter l'école qui, dans une province, ont été vaccinés contre la poliomyélite. Supposons en outre, dans un souci de simplicité, que la province comporte en tout 500 zones d'énumération de même taille, chacune comptant 25 enfants d'âge scolaire. Les zones d'énumération seront les grappes dans cet exemple. Supposons que nous sélectionnions 10 zones d'énumération par échantillon aléatoire simple sans remplacement sur les 500 zones d'énumération de la province et que la proportion d'enfants vaccinés soit mesurée pour chaque zone d'énumération sélectionnée, les résultats étant ceux indiqués au tableau 7.6 ci-après.

Tableau 7.6
Proportions d'enfants en âge de fréquenter l'école qui ont été vaccinés dans 10 zones d'énumération

Zone d'énumération sélectionnée ($\hat{P}_i$)	1	2	3	4	5	6	7	8	9	10
Proportion de l'échantillon ($\hat{P}_i$)	$\frac{8}{25}$	$\frac{10}{25}$	$\frac{12}{25}$	$\frac{14}{25}$	$\frac{15}{25}$	$\frac{17}{25}$	$\frac{20}{25}$	$\frac{20}{25}$	$\frac{21}{25}$	$\frac{23}{25}$

Dans cet exemple, l'estimation de la proportion d'enfants vaccinés dans la province est :

$$\hat{P} = \frac{160}{250} = 0{,}64, \text{ ou } 64\ \%.$$

En outre, la variance de l'échantillon est :

$$s_p^2 = \frac{1}{10-1}\sum_{i=1}^{10}(\hat{P}_i - \hat{P})^2 = 0{,}040533.$$

Par conséquent, la variance de la proportion estimative est :

$$v(\hat{P}) = (1 - \frac{10}{500}) \times \frac{0{,}040533}{10} = \ 0{,}003972.$$

Il y a lieu de noter que, dans le cas d'un échantillonnage aléatoire simple, la variance estimative de la proportion estimative est :

$$v(\hat{P}_{SRS}) = (1 - \frac{250}{12\ 500}) \times \frac{0{,}64\ (1-0{,}64)}{250-1} = \ 0{,}0009078.$$

Par conséquent, l'effet de conception pour cette conception d'échantillonnage en grappes est de $\frac{0{,}003972}{0{,}0009078} = 4{,}38$

et la taille effective de l'échantillon est de $\frac{250}{4{,}38} = 57$.

Cela signifie que l'estimation fondée sur la grappe de 250 unités a la même variance que celle qui est fondée sur un échantillon aléatoire simple de 57 unités.

7.5. Caractéristiques communes des conceptions d'échantillonnage et des données d'enquêtes sur les ménages

7.5.1. Écart des conceptions des enquêtes sur les ménages par rapport à l'échantillonnage aléatoire simple

33. Comme indiqué ci-dessus, la méthode de l'échantillonnage aléatoire simple est rarement employée dans la pratique pour des enquêtes de grande envergure sur les ménages car son utilisation est trop onéreuse. Cependant, il importe de bien comprendre cette conception car elle constitue le fondement théorique de conceptions plus complexes. La plupart des conceptions d'échantillonnage utilisées dans le contexte d'enquêtes sur les ménages s'écartent de l'échantillonnage aléatoire simple en raison de la présence d'une ou de plusieurs des trois caractéristiques ci-après :

a) Stratification à une ou plusieurs étapes de l'échantillonnage;

b) Mise en grappes des unités lors d'une ou plusieurs étapes de l'échantillonnage, ce qui réduit les coûts mais gonfle la variance des estimations en raison des corrélations qui existent entre les unités de la même grappe;

c) Pondération visant à compenser des imperfections de l'échantillon, comme des probabilités inégales de sélection, la non-réponse et la non-couverture (voir le chapitre 6 pour de plus amples détails).

34. Une conception est appelée *complexe* si elle présente une ou plusieurs des caractéristiques susmentionnées. La plupart des conceptions d'enquêtes sur les ménages sont complexes et ne répondent donc pas aux hypothèses qui sont à la base d'un échantillonnage aléatoire simple. Par conséquent, analyser les données d'enquêtes sur les ménages comme si elles étaient générées au moyen d'un échantillonnage aléatoire simple entraînerait des erreurs dans l'analyse et dans les déductions fondées sur ces données. En outre, comme on l'a déjà dit, les estimations qui présentent de l'intérêt dans le cas de la plupart des enquêtes sur les ménages ne peuvent pas être exprimées sous forme de fonctions linéaires des observations, de sorte qu'il peut ne pas y avoir de formule fixe pour estimer les variances. Les sections ci-après analysent la question des méthodes d'estimation de la variance pour les conceptions d'enquêtes sur les ménages qui tiennent compte des complexités esquissées ci-dessus.

7.5.2. Préparation des fichiers de données aux fins de l'analyse

35. Les données rassemblées au cours des enquêtes menées dans les pays en développement ne se prêtent parfois pas à une analyse allant au-delà de fréquences et de tabulations de base. Il y a à cela plusieurs raisons. Premièrement, il se peut que la documentation technique concernant la conception de l'échantillon utilisé soit très limitée, voire inexistante. Deuxièmement, il arrive que les fichiers de données n'aient pas le format, la structure et les informations requis pour permettre une analyse poussée. Troisièmement, il se peut que les logiciels appropriés et les compétences techniques voulues fassent défaut.

36. Si l'on veut pouvoir analyser comme il convient les données des enquêtes pas sondage, la base de données connexe doit contenir toutes les informations reflétant le processus de sélection de l'échantillon (voir le chapitre 5 pour de plus amples détails). En particulier, la base de données doit comporter des étiquettes appropriées pour les strates prévues par la conception de l'échantillon, les unités primaires d'échantillonnage (UPE), les unités secondaires d'échantillonnage (USE), etc. Parfois, les données et les UPE effectivement utilisées pour la sélection de l'échantillon doivent être modifiées afin d'estimer la variance. Ces modifications sont nécessaires pour que la conception effective de l'échantillon corresponde à l'une des options disponibles au moins des logiciels d'analyses statistiques (voir la section 7.9). Les strates et les UPE créées pour estimer la variance sont parfois appelées pseudo-strates ou strates de variance, et pseudo-UPE ou UPE de variance. Les variables pertinentes prévues par la conception de l'échantillon, ainsi que les variables créées pour l'estimation de la variance, doivent être entrées dans la base de données en même temps que la documentation indiquant comment ces variables sont estimées et utilisées. Il faut pour estimer la variance au moins trois variables: la pondération de l'échantillon, la strate (ou pseudo-strate) et l'UPE (ou la pseudo-UPE). Ces trois variables résument la conception de l'échantillon et leur inclusion dans la série de données permet d'analyser comme il convient les informations rassemblées compte tenu des complexités de la conception de l'échantillon.

37. En outre, il faut calculer des pondérations pour chacune des unités d'échantillonnage figurant dans le fichier de données. Ces pondérations doivent refléter la probabilité de sélection de chaque unité d'échantillonnage et compenser la non-réponse et d'autres déficiences de l'échantillon. Les pondérations et les étiquettes utilisées pour identifier les variables de conception sont nécessaires pour pouvoir estimer comme il convient la variabilité et les estimations issues de l'enquête. Comme mentionné au chapitre 6 et dans les sections précédentes du présent chapitre, les pondérations d'échantillonnage sont importantes non seulement pour pouvoir générer des estimations d'enquête appropriées mais aussi pour évaluer les erreurs d'échantillonnage qui caractérisent ces estimations. Il est donc essentiel que toutes les informations concernant les pondérations figurent dans le fichier de données. En particulier, en présence d'une non-réponse, d'une post-stratification ou d'autres types d'ajustement, la documentation doit contenir une description des procédures suivies pour procéder à ces ajustements.

7.5.3. Types d'estimations d'enquête

38. Pour la plupart des enquêtes sur les ménages, les estimations les plus communément recherchées se présentent sous forme de totaux et de ratios. Prenons le cas d'une conception stratifiée en

trois étapes, avec une sélection d'UPE à la première étape, d'USE à la deuxième étape et de ménages à la troisième étape. L'estimation d'enquête d'un total peut être exprimée comme :

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} \sum_{k=1}^{l_j} W_{hijk} Y_{hijk} \qquad (7.11)$$

où W_{hijk} est la pondération finale du $k^{ième}$ ménage (k = 1, ... 1_j) sélectionné dans la $j^{ième}$ USE (j = 1, ... m_i) dans la $i^{ième}$ UPE (i = 1, ... n_h) dans la $h^{ième}$ strate (h = 1, ... H); et Y_{hijk} est la valeur de la variable Y pour le $k^{ième}$ ménage sélectionné dans la $j^{ième}$ USE dans la $i^{ième}$ UPE dans la $h^{ième}$ strate.

39. Au niveau le plus élémentaire, les pondérations associées aux unités d'échantillonnage sont inversement proportionnelles aux probabilités de sélection des unités faisant partie de l'échantillon. Toutefois, des méthodes plus pointues sont fréquemment utilisées pour calculer les pondérations à appliquer lors de l'analyse. Certaines de ces méthodes sont décrites au chapitre 6 et dans les références qui y sont citées.

40. L'estimation d'enquête d'un ratio est définie comme étant :

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \qquad (7.12)$$

où $\hat{Y}$ et $\hat{X}$ sont des estimations des totaux pour les variables Y et X respectivement, calculées comme indiqué dans l'équation (7.12) ci-dessus.

41. Dans un échantillonnage à plusieurs étapes, les moyennes et les proportions ne sont que des cas particuliers de l'estimation d'un ratio. Dans le cas des moyennes, la variable X, dans le dénominateur du ratio, est une variable de dénombrement définie comme étant égale à l'unité pour chaque élément, de sorte que le dénominateur soit la somme des pondérations. Dans le cas d'une proportion, la variable X, dans le dénominateur, est également définie comme étant égale à l'unité pour tous les éléments, et la variable Y, dans le numérateur, est une variable binaire définie comme égale à 0 ou à 1, selon que l'unité observée possède ou non la caractéristique dont on cherche à estimer la proportion. Dans le cas de la plupart des enquêtes sur les ménages, le dénominateur dans l'estimation du ratio est défini comme la population totale, le nombre total de femmes, le nombre total d'hommes, la population rurale totale, la population totale d'une province ou d'un district donné, etc.

7.6. Lignes directrices concernant la présentation des informations relatives aux erreurs d'échantillonnage

7.6.1. Informations à fournir

42. Dans le cas d'enquêtes nationales de grande envergure comportant un grand nombre de variables et de domaines et plusieurs objectifs fréquemment concurrents, il n'est pas possible de présenter chaque estimation sans exception avec l'erreur d'échantillonnage connexe. Cela risquerait non seulement d'accroître énormément le volume de la publication mais aussi d'encombrer la présentation des résultats proprement dits. Étant donné la variabilité prévisible des estimations des erreurs d'échantillonnage elles-mêmes, présenter les résultats pour un trop grand nombre de variables individuelles pourrait susciter une confusion et donner l'impression que, d'une façon générale, la qualité des don-

nées rassemblées est inégale. Il est beaucoup plus utile de ne présenter des informations touchant les erreurs d'échantillonnage que pour quelques-unes des caractéristiques les plus importantes, en laissant le reste pour un appendice.

43. Lorsque l'on présente des informations concernant les erreurs d'échantillonnage, il importe de ne pas perdre de vue leur impact potentiel sur l'interprétation des résultats de l'enquête et sur les décisions que les décideurs pourraient être appelés à prendre sur la base de cette interprétation. Les erreurs d'échantillonnage ont toujours été considérées comme une composante seulement de l'ensemble des erreurs que présente l'enquête, et pas toujours les plus importantes. Dans le cas de certaines enquêtes, les erreurs autres que d'échantillonnage (voir le chapitre 8) peuvent avoir un impact plus marqué que les erreurs d'échantillonnage sur la qualité des données d'enquête dans leur ensemble. Il est donc recommandé que les informations concernant les erreurs d'échantillonnage indiquent les principales sources d'erreurs autres que d'échantillonnage ainsi que certaines évaluations qualitatives de leur impact sur la qualité globale des données d'enquête. Comme les erreurs d'échantillonnage revêtent une importance plus critique aux niveaux de décomposition moins élevés, il est également recommandé d'inclure une mise en garde concernant la mesure dans laquelle les données d'enquête peuvent être décomposées.

44. D'une manière générale, les informations concernant les erreurs d'échantillonnage doivent être suffisamment détaillées pour faciliter une interprétation correcte des résultats de l'enquête et répondre aux besoins de toute la gamme d'usagers des données, qu'il s'agisse du grand public ou des décideurs, qui se fonderont sur les résultats de l'enquête pour formuler des politiques, de l'analyste, qui étudie à fond les données et rend compte des résultats, ou du statisticien, qui s'intéresse surtout à l'efficience statistique de la conception retenue en comparaison d'autres options possibles ainsi qu'aux caractéristiques de cette conception qui pourraient être utilisées pour de futures enquêtes.

7.6.2. Comment présenter les informations sur les erreurs d'échantillonnage

45. Les erreurs d'échantillonnage peuvent être présentées sous trois formes différentes :

a) Valeurs absolues d'erreurs types;

b) Erreurs types relatives (racine carrée de variances relatives);

c) Intervalles de confiance.

46. Le choix entre ces trois formes de présentation dépend de la nature de l'estimation. Lorsque les estimations varient pour ce qui est de leur taille et des unités de mesure, la même valeur d'erreurs types peut être applicable aux estimations lorsqu'elles sont exprimées en termes relatifs; il serait donc plus efficace de présenter l'erreur type relative. D'une façon générale, cependant, les erreurs types absolues sont beaucoup plus faciles à comprendre et à rattacher à l'estimation, surtout dans le cas de pourcentages, de proportions et de taux. Pour utiliser des intervalles de confiance, il faut choisir le niveau de confiance (par exemple 90, 95 ou 99 %). Comme ce niveau varie selon les objectifs de l'enquête et la précision que doivent avoir les estimations, il est important de spécifier le niveau de confiance utilisé dans la présentation des informations concernant les erreurs d'échantillonnage et de conserver ensuite ce niveau de confiance tout au long de la présentation des résultats afin de pouvoir en déterminer la signification. Comme indiqué précédemment, l'intervalle le plus fréquemment utilisé dans la pratique est l'intervalle de confiance de 95 % (voir les paragraphes 30 et 22 des chapitres 3 et 7 respectivement), c'est-à-dire :

$$\text{Estimation} \pm 1{,}96 \quad \text{Erreur type} \tag{7.13}$$

47. Pour de plus amples détails sur la présentation des informations concernant les erreurs d'échantillonnage, y compris les lignes directrices spécifiques à suivre pour différentes catégories d'usagers, et pour un certain nombre d'exemples, voir Organisation desNations Unies (1993) et les références qui y sont citées.

7.6.3. Règles approximatives concernant les informations à fournir au sujet des erreurs types

48. Une règle approximative fréquemment utilisée consiste à indiquer l'erreur type au niveau des deux chiffres les plus significatifs et de signaler ensuite l'estimation ponctuelle correspondante avec le même nombre de décimales que l'erreur type. Par exemple :

1. Si l'estimation ponctuelle est de 73 456 avec une erreur type de 2 345, nous signalerons l'estimation ponctuelle comme étant de 73 500 et l'erreur type de 2 300.

2. Si l'estimation ponctuelle est de 1,54328 avec une erreur type de 0,01356, nous signalerons l'estimation ponctuelle comme étant de 1,543 avec une erreur type de 0,014.

49. Le raisonnement général qui est à la base de cette règle paraît être lié à la t-statistique. La présence de deux chiffres significatifs dans l'erreur type et d'un nombre correspondant de chiffres dans l'estimation ponctuelle permet de faire en sorte que l'erreur due à l'arrondissement des chiffres n'affecte pas trop la t-statistique connexe, tout en évitant de donner l'impression d'une précision excessive en présentant les estimations ponctuelles avec un grand nombre de chiffres dépourvus de pertinence. Il y a lieu de noter toutefois que cette règle n'est pas nécessairement applicable lorsque la t-statistique ne présente pas d'intérêt primordial.

7.7. Méthodes d'estimation de la variance dans le contexte des enquêtes sur les ménages

50. Nous décrirons brièvement dans cette section certaines méthodes classiques d'estimation des variances ou des erreurs d'échantillonnage pour les estimations fondées sur les données d'enquête. Les méthodes d'estimation des erreurs d'échantillonnage peuvent être classées en quatre grandes catégories :

a) Méthodes exactes;

b) Estimation de la variance de la grappe ultime;

c) Approximations par linéarisation;

d) Méthodes de réplication.

Nous discuterons maintenant brièvement, tour à tour, de chacune de ces méthodes. Le lecteur intéressé pourra trouver de plus amples détails dans des ouvrages comme Kish et Frankel (1974), Wolter (1985) et Lehtonen et Pahkinen (1995).

7.7.1. Méthodes exactes

51. Les sections 7.2 et 7.4 ont donné plusieurs exemples de méthodes exactes d'estimation de la variance pour des conceptions d'échantillonnage type. Ces méthodes constituent la meilleure approche de l'estimation de la variance lorsqu'elles peuvent être utilisées. Cependant, leur application pour le calcul des variances d'échantillonnage des estimations fondées sur les données provenant d'enquêtes sur les ménages est compliquée par plusieurs facteurs. Premièrement, les conceptions utilisées pour la plupart des enquêtes sur les ménages sont plus complexes que l'échantillonnage aléatoire simple (voir la section 7.5.1 ci-dessus). Deuxièmement, les estimations à étudier peuvent ne pas se présenter sous forme de fonctions linéaires simples des valeurs observées, de sorte que, fréquemment, la variance d'échantillonnage peut ne pas être exprimée par une formule toute faite comme celle qui concerne la moyenne de l'échantillon dans le cas d'un échantillonnage aléatoire simple ou d'un échantillonnage stratifié. En outre, l'application des méthodes exactes dépend de la conception dont il s'agit, de l'estimation à étudier et des procédures de pondération utilisées.

52. Nous discuterons dans les sections ci-après des méthodes d'estimation de la variance pour les conceptions d'échantillonnage habituellement employées pour les enquêtes sur les ménages. Ces méthodes sont conçues de manière à surmonter les défaillances qui caractérisent les méthodes exactes.

7.7.2. Méthode d'estimation de la variance de la grappe ultime

53. La méthode d'estimation de la variance de la grappe ultime (voir Hansen, Hurwitz et Madow, 1953, p. 257-259) peut être utilisée pour estimer les variances des estimations sur la base d'un échantillon généré par une conception d'échantillonnage complexe. Selon cette méthode, la grappe ultime comprend l'intégralité de l'échantillon tiré d'une UPE, quel que soit l'échantillonnage réalisé lors des étapes suivantes. Les estimations de la variance sont calculées en utilisant uniquement les totaux au niveau des UPE, sans devoir calculer les composantes de la variance à chacune des étapes de la sélection.

54. Supposons qu'un échantillon de n_h UPE soit sélectionné dans la strate h (quel que soit le nombre d'étapes à l'intérieur des UPE). L'estimation du total de la strate h est alors donnée par :

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} \qquad (7.14)$$

$$\text{où } \hat{Y}_{hi} = \sum_{j=1}^{m_i} W_{hijk}\, Y_{hijk}$$

Il y a lieu de noter que l'estimation $\hat{Y}_{hi}$ au niveau de l'UPE est une estimation de $\frac{\hat{Y}_h}{n_h}$. Ainsi, la variance des estimations au niveau des UPE est donnée par :

$$v(\hat{Y}_{hi}) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \qquad (7.15)$$

et la variance de leur total, $\hat{Y}_{hi}$, le total au niveau de la strate, estimée à partir d'un échantillon aléatoire de taille n_h représentant le total de la population pour la strate h, est donnée par :

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \qquad (7.16)$$

55. Il convient de noter qu'une manipulation algébrique simple donne l'expression équivalente ci-dessous pour l'estimation de la variance du total de la population pour la strate *h* :

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1}\left\{\sum_{i=1}^{n_h}\hat{Y}_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h}\hat{Y}_{hi}\right)^2}{n_h}\right\} \tag{7.17}$$

56. Enfin, avec un échantillonnage indépendant dans les différentes strates, l'estimation de la variance pour le total global de la population est obtenue en prenant la somme des variances des totaux au niveau des strates, c'est-à-dire :

$$v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h) \tag{7.18}$$

Parfois, un facteur de correction de la population finie $(1 - n_h/N_h)$ est utilisé dans les formules susmentionnées.

57. L'équation (7.18) est remarquable en ce sens que la variance du total estimatif est une fonction des totaux dûment pondérés des UPE $\hat{Y}_{hi}$ seulement, sans aucune référence à la structure et à la qualité de l'échantillonnage à l'intérieur des UPE. Cela simplifie considérablement la formule d'estimation de la variance étant donné qu'il n'est pas nécessaire de calculer les composantes de la variance imputables aux autres étapes de l'échantillonnage à l'intérieur des UPE. Ainsi, la méthode d'évaluation de la variance de la grappe ultime est plus souple et peut être utilisée pour différentes conceptions, ce qui est effectivement l'un des principaux avantages de cette méthode et l'une des principales raisons pour lesquelles elle est largement utilisée dans les enquêtes.

58. L'estimation de la variance du ratio $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ est donnée par :

$$v(\hat{R}) = \frac{1}{\hat{X}^2}\left\{v(\hat{Y}) + \hat{R}^2 v(\hat{X}) - 2\,\mathrm{cov}(\hat{Y},\ \hat{X})\right\} \tag{7.19}$$

où $v(\hat{Y})$ et $v(\hat{X})$ sont calculées selon les formules d'estimation de la variance d'un total estimatif; et

$$\mathrm{cov}(\hat{Y}, \hat{X}) = \sum_{h=1}^{H}\left\{\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}\left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h}\right)\left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h}\right)\right\} \tag{7.20}$$

ou, ce qui est la même chose,

$$\mathrm{cov}(\hat{Y}, \hat{X}) = \frac{n_h}{n_h - 1}\left\{\sum_{i=1}^{n_h}\hat{X}_{hi}\,\hat{Y}_{hi} - \frac{\left(\sum_{i=1}^{n_h}\hat{X}_{hi}\right)\left(\sum_{i=1}^{n_h}\hat{Y}_{hi}\right)}{n_h}\right\} \tag{7.21}$$

59. Cette formule de calcul de la variance d'un ratio peut être simplifiée étant donné que la variance relative du ratio est approximativement égale à la différence entre les variances relatives du numérateur et du dénominateur. L'on se souviendra que la variance relative de l'estimation est le ratio entre sa variance et son carré. Ainsi, pour un ratio estimatif $\hat{R}$, la variance relative, dénotée par $relvar\,(\hat{R})$, est donnée par :

$$relvar\,(\hat{R}) = \frac{v(\hat{R})}{\hat{R}^2} \tag{7.22}$$

Par conséquent, l'estimation de la variance du ratio est donnée par la formule

$$v\,(\hat{R}) = \hat{R}^2\, relvar\,(\hat{R}) = \hat{R}^2 \left\{ relvar\,(\hat{Y}) - relvar\,(\hat{X}) \right\} \tag{7.23}$$

60. La méthode d'évaluation de la variance de la grappe ultime utilisée pour le calcul des erreurs d'échantillonnage de totaux estimatifs et de ratios peut être schématisée comme suit :

- *Étape 1.* Pour chaque strate séparément, calculer l'estimation pondérée $\hat{Y}_{hi}$ pour la caractéristique étudiée, Y, pour chaque UPE (conformément aux procédures de pondération spécifiées au chapitre 6).
- *Étape 2.* Calculer le carré de la valeur estimative calculée pour chaque UPE lors de l'étape 1.
- *Étape 3.* Calculer la somme des valeurs de l'étape 2 pour toutes les UPE de la strate.
- *Étape 4.* Calculer la somme des totaux estimatifs des UPE de l'étape 1 pour toutes les UPE.
- *Étape 5.* Porter au carré le résultat de l'étape 4 divisé par n_h, nombre d'UPE de la strate.
- *Étape 6.* Soustraire le résultat de l'étape 5 de celui de l'étape 3 et multiplier cette différence par le facteur $n_h/(n_h - 1)$, qui est la variance estimative de la caractéristique étudiée au niveau de la strate.
- *Étape 7.* Additionner le résultat de l'étape 6 pour toutes les strates pour obtenir la variance estimative globale de la caractéristique étudiée.
- *Étape 8.* Calculer la racine carrée du résultat de l'étape 7 pour obtenir l'erreur d'échantillonnage estimative pour la caractéristique étudiée.

61. Pour calculer l'erreur d'échantillonnage estimative pour les ratios, comme des proportions estimatives, nous procéderons comme suit :

- *Étape 9.* Calculer la variance relative du numérateur, $\hat{Y}$, en divisant le résultat de l'étape 7 par le carré de l'estimation du numérateur.
- *Étape 10.* Répéter l'étape 9 pour obtenir la variance relative du dénominateur, $\hat{X}$.
- *Étape 11.* Soustraire le résultat de l'étape 10 de celui de l'étape 9.
- *Étape 12.* Multiplier le résultat de l'étape 11 par le carré du ratio estimatif, $\hat{R}$, qui est la variance estimative de $\hat{R}$.
- *Étape 13.* Calculer la racine carrée du résultat de l'étape 12 pour obtenir l'erreur d'échantillonnage estimative pour $\hat{R}$.

Exemple 4

Nous prendrons maintenant un exemple hypothétique pour illustrer la méthode d'estimation de la variance de la grappe ultime. Supposons que nous souhaitions estimer les dépenses hebdomadaires totales d'alimentation des ménages de la ville A. Nous envisageons de mener une enquête reposant sur une conception stratifiée en grappes à trois étapes comportant trois strates, deux UPE devant être sélectionnées dans chaque strate et deux ménages devant être sélectionnés dans chaque UPE retenue. Les dépenses hebdomadaires d'alimentation sont alors consignées pour chaque ménage interrogé. Le tableau 7.7 présente les données tirées de l'enquête, y compris les pondérations (W_{hij}) et les dépenses hebdomadaires d'alimentation en dollars (Y_{hij}), pour chaque ménage sélectionné.

Tableau 7.7
Dépenses hebdomadaires d'alimentation des ménages, par strate

Strate	UPE	Ménage	Pondération (W_{hij})	Dépenses en dollars (Y_{hij})	Total W_{hij}*Y_{hij}
1	1	1	1	30	30
		2	1	28	28
	2	1	3	12	36
		2	3	15	45
2	1	1	5	6	30
		2	5	7	35
	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
	2	1	4	13	52
		2	4	15	60
Total			42		474

62. Conformément aux formules indiquées dans le chapitre 6, une estimation du total des dépenses hebdomadaires d'alimentation des ménages de la ville A est donnée par :

$$\hat{Y} = \sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}\, Y_{hij} = 474.$$

En outre, une estimation des dépenses hebdomadaires moyennes d'alimentation des ménages est donnée par :

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}\, Y_{hij}}{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}} = \frac{474}{42} = 11 \text{ (chiffre arrondi au dollar le plus proche).}$$

63. Nous allons maintenant suivre les étapes prévues par la méthode d'estimation de la variance de la grappe ultime pour les appliquer aux colonnes du tableau 7.8 ci-après. Les numéros de colonnes correspondent aux étapes indiquées ci-dessus.

Tableau 7.8
Application des étapes de la méthode d'estimation de la variance de la grappe ultime

Strate	UPE	Étape 1	Étape 2	Étape 3	Étape 4	Étape 5	Étape 6	Étape 7
1	1	58	3 364	9 925	139	9 660,5	529	
	2	81	6 561	_	-		-	
2	1	65	4 225	8 849	133	8 844,5	9	
	2	68	4 624	_	-		-	
3	1	90	8 100	20 644	202	20 402	484	
	2	112	12 544	-	-		-	
Total								1 022

64. Les estimations de la variance au niveau des strates sont de 529 pour la strate 1, de 9 pour la strate 2 et de 484 pour la strate 3. L'estimation globale de la variance de l'estimation du revenu hebdomadaire total des ménages (étape 7 de notre exemple) est obtenue en faisant la somme des estimations au niveau des strates, soit 1 022.

7.7.3. Approximations par linéarisation

65. La plupart des estimations que l'on cherche à établir au moyen d'enquêtes sur les ménages sont non linéaires. L'on peut en citer comme exemple l'indice de la masse corporelle moyenne des enfants en âge de fréquenter l'école, la proportion du revenu consacrée au logement dans une ville déterminée, le ratio entre la probabilité qu'un sous-groupe de population possède une caractéristique donnée et celle qu'un autre sous-groupe la possède également, etc. Selon la méthode de linéarisation, ces estimations non linéaires sont « linéarisées » conformément à la méthode de linéarisation par série de Taylor. Cette méthode consiste à exprimer l'estimation en termes d'une expansion par série de Taylor puis à calculer de façon approximative la variance de l'estimation en se référant à la variance de premier ordre ou à la partie linéaire de cette expansion en utilisant les méthodes exactes présentées dans les sections précédentes.

66. Supposons que nous souhaitions estimer la variance d'une estimation z d'un paramètre $\underline{Z}$ et que z soit une fonction non linéaire d'une estimation simple $y_1, y_2, \ldots, y_m$ des paramètres $Y_1, Y_2, \ldots, Y_m$, c'est-à-dire :

$$z = f(y_1, y_2, \ldots \ldots, y_m) \tag{7.24}$$

À supposer que z soit proche de Z, l'expansion par série de Taylor de z au premier degré dans $z - \underline{Z}$ est :

$$z = Z + \sum_{i=1}^{m} d_i (y_i - Y_i) \tag{7.25}$$

où d_i' est le dérivé partiel de z pour y_i', c'est-à-dire : $d_i = \dfrac{\partial z}{\partial y_i}$,

qui est une fonction de l'estimation de base y_i. Cela signifie que la variance de $\underline{z}$ peut être calculée de façon approximative par la variance de la fonction linéaire dans l'équation (7.24) ci-dessus, que

nous savons calculer au moyen des méthodes exactes présentées dans les sections précédentes, c'est-à-dire :

$$v(z) = v(\sum d_i y_i) = \sum_{i=1}^{m} d_i^2 v(y_i) + \sum_{i \neq j} d_i d_j \operatorname{cov}(y_i, y_j) \tag{7.26}$$

67. L'équation (7.26) fait intervenir une matrice de covariance $m \times m$ de m estimations de base $y_1, y_2, \ldots, y_m$, avec des termes de variance m et des termes identiques de covariance $m(m - 1)/2$, qui peuvent être évalués au moyen des méthodes exactes de statistiques linéaires présentées dans la section précédente.

Exemple 5 (variance d'un ratio)

Pour illustrer la méthode de linéarisation, prenons l'estimation de la variance pour le ratio :

$$z = r = \frac{y}{x} \tag{7.27}$$

Il y a lieu de noter que, dans ce cas, $\frac{\partial r}{\partial y} = \frac{1}{x}$ et $\frac{\partial r}{\partial x} = -\frac{y}{x^2} = -\frac{r}{x}$. Par conséquent,

$$v(r) = \frac{1}{x^2}\left\{v(y) + r^2 v(x) - 2r \operatorname{cov}(y, x)\right\}, \tag{7.28}$$

qui est l'expression familière de la variance d'un ratio trouvée dans la plupart des manuels d'échantillonnage.

68. La linéarisation est largement utilisée dans la pratique car elle peut être appliquée à presque toutes les conceptions d'échantillon et à toute statistique qui peut être linéarisée, c'est-à-dire exprimée sous forme de fonction statistique linéaire familière comme des moyennes ou des totaux, avec des coefficients provenant de dérivés partiels, comme l'exige la méthode de linéarisation par série de Taylor. Une fois linéarisée, la variance de l'estimation non linéaire peut être calculée de façon estimative au moyen des méthodes exactes décrites ci-dessus [voir Cochran (1977) et Lohr (1999) pour des informations techniques concernant le processus de linéarisation, avec des exemples].

7.7.3.1. Avantages

69. Comme la méthode de linéarisation d'estimation de la variance est utilisée depuis longtemps, sa théorie est bien développée et elle peut être appliquée à un plus grand nombre de conception d'échantillonnage que celles auxquelles peuvent être appliquées les méthodes de réplication (décrites ci-après). Si les dérivés partiels sont connus et si les termes quadratiques et les termes de rang supérieur dans la linéarisation par série de Taylor sont négligeables, la linéarisation produit une estimation approximative de la variance pour presque tous les estimateurs linéaires étudiés, comme les ratios et les coefficients de régression.

7.7.3.2. Limitations

70. La linéarisation ne donne de bons résultats que si les hypothèses susmentionnées concernant les dérivés partiels et les termes de rang supérieur sont correctes. Autrement, de graves distorsions

peuvent être introduites dans les estimations. En outre, il est généralement difficile d'appliquer cette méthode à des fonctions complexes comportant des pondérations. Il faut élaborer une formule séparée pour chaque type d'estimateur, ce qui peut exiger une programmation spéciale. Cette méthode ne peut pas être appliquée à des statistiques comme des valeurs médianes et d'autres percentiles qui ne sont pas immédiatement fonctions de totaux ou de moyennes de la population.

71. En outre, il est difficile, par l'approche de linéarisation, d'introduire des ajustements pour tenir compte de la non-réponse et de la non-couverture, car cette méthode dépend de la conception de l'échantillon, de l'estimation étudiée et des procédures de pondération. Il faut également que les informations concernant la conception de l'échantillon (strates, UPE, pondérations) soient incorporées au fichier de données.

7.7.4. Réplication

72. L'approche de la réplication consiste à tirer des données des sous-échantillons répétés, ou *réplicats,* à recalculer pour chaque réplicat et pour l'ensemble de l'échantillon l'estimation pondérée et à calculer ensuite la variance en tant que fonction des écarts des estimations de ces réplicats par rapport aux estimations concernant l'ensemble de l'échantillon. Cette approche peut être résumée par les étapes suivantes :

- *Étape 1.* Supprimer différents sous-échantillons de l'ensemble de l'échantillon pour former des réplicats.
- *Étape 2.* Produire des pondérations des réplicats en répétant le processus d'estimation pour chaque réplicat.
- *Étape 3.* Produire une estimation à partir de l'ensemble de l'échantillon puis de chaque série de pondération des réplicats.
- *Étape 4.* Calculer la variance de l'estimation comme écarts carrés des estimations des réplicats à partir de l'estimation concernant l'ensemble de l'échantillon.

73. Supposons par exemple qu'il soit créé à partir d'un échantillon k réplicats caractérisés chacun par des estimations $\hat{\theta}_1$, $\hat{\theta}_2$, ... $\hat{\theta}_k$ d'un paramètre $\hat{\theta}$, et supposons en outre que l'estimation fondée sur l'ensemble de l'échantillon soit $\hat{\theta}_0$. Ainsi, l'estimation de la variance fondée sur la réplication est donnée par la formule :

$$\text{var}\,(\hat{\theta}) = \frac{1}{c}\sum_{r=1}^{k}(\hat{\theta}_r - \hat{\theta}_0)^2, \tag{7.29}$$

où c est une constante qui dépend de la méthode d'estimation. Les méthodes de réplication diffèrent par la valeur de la constante et les modalités de formation des réplicats (voir la section 7.7.5 pour un bref aperçu des méthodes de réplication les plus communément utilisées).

7.7.4.1. Structure des fichiers de données

74. Quelle que soit la technique de réplication, la structure des fichiers de données demeure la même, comme le montre le tableau 7.9 ci-après.

Tableau 7.9
Structure des fichiers de données selon l'approche de réplication

Fichier	Données	Pondération de l'ensemble de l'échantillon	Pondérations des réplicats				
			1	2	3	....	*k*
1	Données 1	W_1	W_{11}	0	W_{13}	…	W_{1k}
2	Données 2	W_2	0		W_{23}	…	W_{2k}
3	Données 3	W_3	W_{31}		0	…	W_{3k}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
N	Données *n*	W_n	W_{n1}		W_{n3}	…	0

Note : Les points dénotent des séries.

7.7.4.2. Avantages

75. Le principal avantage de l'approche de réplication par rapport à l'approche de linéarisation est qu'elle utilise essentiellement la même méthode d'estimation, quelle que soit la statistique à estimer (étant donné que l'estimation de la variance est fonction de l'échantillon et non de l'estimation), tandis qu'une approximation par linéarisation doit être établie à la suite d'une analyse de chaque statistique, ce qui peut être laborieux dans le cas d'enquêtes sur les ménages de grande envergure visant un grand nombre de caractéristiques. En outre, les techniques de réplication sont d'utilisation commode et sont applicables à presque toutes les statistiques, quelles soient linéaires ou non linéaires. Avec la méthode de réplication, il est facile de calculer des estimations pour des sous-groupes de population, et les effets des ajustements visant à compenser la non-réponse et d'autres facteurs peuvent être reflétés dans les pondérations des réplicats.

7.7.4.3. Limitations

76. Les techniques de réplication exigent un gros travail sur ordinateur, essentiellement parce qu'il faut calculer une série de pondérations pour les réplicats sélectionnés de sorte que chaque réplicat représente comme il convient la même population que l'ensemble de l'échantillon. En outre, la formation de réplicats peut être compliquée par les contraintes liées à la conception de l'échantillon (voir la section 7.7.5 ci-dessous), ce qui peut parfois conduire à surestimer les erreurs d'échantillonnage.

77. Nous conclurons notre comparaison générale des approches de linéarisation et de réplication de l'estimation des erreurs d'échantillonnage en disant que ces deux approches ne donnent pas des estimations identiques des erreurs d'échantillonnage. Il ressort cependant d'études empiriques (voir Kish et Frankel, 1974) que, pour les échantillons nombreux et de nombreuses statistiques, les différences entre les résultats donnés par ces deux méthodes sont négligeables.

7.7.5. Quelques techniques de réplication

78. Les techniques de réplication les plus communément utilisées sont les suivantes :

a) Groupes aléatoires;

b) Réplication répétée équilibrée (BRR);

c) Réplication « jackknife » (JK1, JK2 et JKn);

d) Méthode « bootstrap ».

Nous examinerons brièvement, tour à tour, chacune de ces techniques.

7.7.5.1. Groupes aléatoires

79. La technique des groupes aléatoires consiste à diviser l'ensemble de l'échantillon en k groupes de manière à préserver la conception de l'enquête; autrement dit, chaque groupe doit représenter une version miniature de l'enquête et refléter la conception de l'échantillon. Par exemple, si l'ensemble de l'échantillon est un échantillon aléatoire simple de taille n, les groupes aléatoires peuvent être constitués en répartissant au hasard les n observations parmi k groupes, chacun de taille n/k. S'il s'agit d'un échantillon en grappes, les UPE sont réparties au hasard parmi les k groupes afin que chaque UPE conserve toutes ses caractéristiques, de sorte que chaque groupe aléatoire demeure un échantillon en grappes. Si l'échantillon est un échantillon stratifié à plusieurs étapes, les groupes aléatoires peuvent être formés en sélectionnant un échantillon d'UPE dans chaque strate. Il y a lieu de noter que le nombre total de groupes aléatoires à former ne peut pas dépasser le nombre d'UPE sélectionnées dans la strate la plus réduite.

80. La méthode des groupes aléatoires peut aisément être utilisée pour estimer les erreurs d'échantillonnage aussi bien pour des statistiques linéaires comme des moyennes et des totaux que pour les fonctions qui en sont directement dérivées et pour des statistiques non linéaires comme des ratios et des percentiles. Il n'est pas nécessaire d'utiliser des logiciels spéciaux pour estimer l'erreur d'échantillonnage, qui est simplement l'écart type des estimations fondées sur les groupes aléatoires constitués à partir de l'ensemble de l'échantillon. Toutefois, il peut être difficile de créer des groupes aléatoires dans le cas de conceptions complexes étant donné que chacun d'eux doit préserver la structure de la conception de l'ensemble de l'enquête. En outre, le nombre de groupes aléatoires peut être limité par la conception de l'enquête elle-même. Par exemple, si la conception prévoit deux UPE par strate, il ne peut être constitué que deux groupes aléatoires et, d'une manière générale, des groupes aléatoires peu nombreux conduisent à des estimations peu précises de l'erreur d'échantillonnage. Une règle approximative générale est qu'il faut avoir au moins dix groupes aléatoires pour obtenir une estimation plus stable de l'erreur d'échantillonnage.

7.7.5.2. Réplication répétée équilibrée

81. La méthode de réplication répétée équilibrée (BRR) suppose une conception prévoyant deux UPE par strate. Pour former un réplicat, l'on divise chaque strate en deux UPE et il est sélectionné l'une d'elles dans chaque strate, selon un schéma prescrit, pour représenter l'ensemble de celle-ci. Cette méthode peut être adaptée à d'autres conceptions en regroupant les UPE en pseudo-strates comportant chacune deux UPE.

7.7.5.3. Réplication « jackknife »

82. Comme la méthode BRR, la réplication « jackknife » est une généralisation des groupes aléatoires qui permet un chevauchement des réplicats. Il y a trois types de réplications « jackknife » : JK1, JK2 et JKn.

83. La technique JK1 est la méthode habituellement utilisée pour supprimer une UPE de l'échantillon aléatoire simple. Toutefois, cette méthode peut être utilisée avec une autre conception si les unités sélectionnées sont groupées en sous-séries aléatoires dont chacune ressemble à l'ensemble de l'échantillon.

84. La méthode JK2 est semblable à la méthode BRR en ce sens qu'elle suppose une conception caractérisée par deux UPE par strate. Dans le cas des UPE autoreprésentées, il peut être créé des paires d'unités secondaires d'échantillonnage (USE). Comme la méthode BRR, la méthode JK2 peut être adaptée à d'autres conceptions en groupant les UPE en pseudo-strates comportant chacune deux UPE. Une UPE est alors éliminée au hasard dans chaque strate, tour à tour, pour constituer les réplicats.

85. La méthode JKn est la méthode de suppression d'une UPE de l'échantillon habituellement utilisée dans le cas des conceptions stratifiées. Pour créer des réplicats, il est supprimé tour à tour une UPE de chaque strate. Les UPE restantes de chaque strate sont repondérées pour estimer le total de la strate. Le nombre de réplicats est égal au nombre d'UPE (ou de pseudo-UPE).

7.7.5.4. Méthode « bootstrap »

86. La méthode « bootstrap » commence par la sélection d'un échantillon reproduisant toutes les caractéristiques les plus importantes de la population dans son ensemble. L'échantillon est alors considéré comme s'il constituait la population dans son ensemble, et il en est tiré des sous-échantillons. Comme précédemment, l'estimation de l'erreur d'échantillonnage est obtenue en tirant l'écart type des estimations calculé sur la base des sous-échantillons de l'écart caractérisant l'échantillon dans son ensemble.

87. La méthode « bootstrap » donne de bons résultats pour les conceptions de caractère général et pour des fonctions qui ne sont pas directement dérivées, comme des percentiles. Cependant, elle exige plus de calculs que les autres méthodes de réplication.

88. Le tableau 7.10 ci-après spécifie la valeur de la constante c de la formule de calcul de la variance [équation (7.28)] qui correspond aux différentes méthodes de réplication.

Tableau 7.10

Valeurs de la constante dans la formule de calcul de la variance pour différentes techniques de réplication

Technique de réplication	Valeur de la constante c dans l'équation (7.28)
Groupe aléatoire	$k(k-1)$
BRR	k
JK1	1
JK2	2
JKn	$k/(k-1)$
Bootstrap	$k-1$

Exemple 6 (jackknife)

Nous donnerons maintenant un exemple numérique simple d'application de la méthode « jackknife » de l'estimation de la variance. Supposons que nous ayons un échantillon de taille 3. Nous pouvons créer trois sous-échantillons de taille 2 en supprimant tour à tour une UPE de l'ensemble de l'échantillon. Le tableau 7.11 ci-dessous présente les valeurs d'une variable (Y). Pour les trois sous-échantillons, un « X » indique quelles sont les UPE faisant partie du sous-échantillon.

Tableau 7.11
Application de la méthode « jackknife » de l'estimation de la variance à un échantillon restreint et à ses sous-échantillons

Unité d'échantillonnage	Variable (Y)	Sous-échantillon (*g*)		
	Y	1	2	3
1	5	X	X	
2	7	X		X
3	9		X	X
Total pour l'échantillon	21			
Moyenne pour l'échantillon	7	6	7	8

Variance de l'échantillon $s^2 = \frac{(5-7)^2+(7-7)^2+(9-7)^2}{3-1} = 4.$

Estimation de la variance de la moyenne de l'échantillon (en ignorant *fpc*) : $\frac{s^2}{n} = \frac{4}{3}.$

Moyenne des moyennes des sous-échantillons : $\frac{\bar{y}_1+\bar{y}_2+\bar{y}_3}{3} = \frac{6+7+8}{3} = 7.$

L'estimation de la variance de la moyenne de l'échantillon selon la méthode « jackknife » est donnée par la formule :

$$V_J(\bar{y}) = \frac{n-1}{n}\sum_{g=1}^{3}(\bar{y}_g - \bar{y})^2 = \frac{3-1}{3}\left[(6-7)^2+(7-7)^2+(8-7)^2\right] = \frac{4}{3},$$

qui est exactement identique à la variance estimative de la moyenne de l'échantillon calculée ci-dessus.

Exemple 7 (formation de réplicats)

Le tableau 7.12 utilise les données de l'exemple 4 (section 7.7.2) pour illustrer la formation de réplicats selon différentes méthodes de réplication ainsi que le calcul des variances par la méthode « jackknife ».

Tableau 7.12
Ensemble de l'échantillon : dépenses par strate

Strate	UPE	Ménage	Pondération W_{hij}	Dépenses Y_{hij}	Total W_{hij}*Y_{hij}
1	1	1	1	30	30
1	1	2	1	28	28
1	2	1	3	12	36
1	2	2	3	15	45
2	1	1	5	6	30
2	1	2	5	7	35
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
3	2	1	4	13	52
3	2	2	4	15	60
Total			42		474

La moyenne estimative fondée sur l'ensemble de l'échantillon est égale à $= \hat{\bar{Y}}_0 = \frac{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}\, Y_{hij}}{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}} = \frac{474}{42} = 11.$

Tableau 7.13
Méthode « jackknife » (*élimination de l'UPE 2 de la strate 1*)

Strate	UPE	Ménage	Pondération (W_{hij})	Dépenses (Y_{hij})	Total W_{hij}*Y_{hij}
1	1	1	1	30	30
1	1	2	1	28	28
2	1	1	5	6	30
2	1	2	5	7	35
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
3	2	1	4	13	52
3	2	2	4	15	60
Total			36		393

La moyenne estimative, sur la base du réplicat ci-dessus, est égale à $= \hat{\bar{Y}}_1 = \frac{393}{36} = 11.$

Ce processus peut être poursuivi en éliminant une UPE à la fois de chaque strate. L'on peut ainsi constituer en tout six réplicats. Le tableau 7.14 ci-après illustre les estimations du revenu hebdomadaire moyen des ménages sur la base de chacun des six réplicats.

Tableau 7.14
Estimations fondées sur les réplicats

Réplicat *j*	UPE supprimée	Estimation $\hat{Y}_j$	$\hat{Y}_j - \hat{Y}_0$	$(\hat{Y}_j - \hat{Y}_0)^2$
1	UPE 2, strate 1	11	0	0
2	UPE 1, strate 1	10	–1	1
3	UPE 2, strate 2	10	–1	1
4	UPE 1, strate 2	12	1	1
5	UPE 2, strate 3	12	1	1
6	UPE 1, strate 3	13	2	4
Total				8

L'estimation par la méthode « jackknife » de la variance de la moyenne estimative est donnée par la formule :

$$\text{var}_{JK}(\hat{\bar{Y}}) = \sum_{h=1}^{H}\left\{\frac{n_h - 1}{n_h}\sum_{j=1}^{n_h}(\hat{\bar{Y}}_j - \hat{\bar{Y}}_0)^2\right\} = \frac{1}{2}\times 8 = 4.$$

(Il y a lieu de noter que, dans cet exemple, H = 3 et $n_h = 2$ pour tous les h.)

89. Nous achèverons cette section en donnant un autre exemple de la formation de réplicats en utilisant la méthode de la réplication répétée équilibrée. Les résultats indiqués au tableau 7.15 ci-dessous correspondent au schéma de suppressions des UPE spécifié dans le titre du tableau.

Tableau 7.15
Méthode de la réplication répétée équilibrée

(Élimination de l'UPE 2 dans les strates 1 et 3 et de l'UPE 1 dans la strate 2)

Strate	UPE	Ménage	Pondération W_{hij}	Dépenses Y_{hij}	Total $W_{hij}*Y_{hij}$
1	1	1	1	30	30
1	1	2	1	28	28
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
Total			18		216

La moyenne estimative fondée sur l'échantillon BRR ci-dessus est égale à $= \hat{\bar{Y}}_{1,BRR} = \frac{216}{18} = 12.$

7.8. Inconvénients de l'utilisation de logiciels statistiques standard pour l'analyse des données d'enquête sur les ménages

90. Pour pouvoir analyser comme il convient les données provenant d'enquêtes sur les ménages, il faut que les erreurs d'échantillonnage qui affectent les estimations soient calculées de manière à tenir compte de la complexité de la conception : stratification, mise en grappes, échantillonnage sur la base

de probabilités inégales, non-réponse et autres ajustements des pondérations (voir le chapitre 6 pour plus amples détails sur le calcul et l'ajustement des pondérations). Les logiciels statistiques standard ne tiennent pas compte de ces complexités car, actuellement, ils tiennent pour acquis que les éléments qui font partie de l'échantillon ont été sélectionnés parmi la population par échantillonnage aléatoire simple. Comme on l'a démontré au chapitre 6, des estimations ponctuelles des paramètres de la population dépendent de la pondération associée à chaque observation. Ces pondérations dépendent des probabilités de sélection des autres caractéristiques de la conception de l'enquête, comme la stratification et la mise en grappes. Lorsqu'ils ne tiennent pas compte des pondérations d'échantillonnage, les logiciels standard produisent des estimations ponctuelles qui sont faussées. Réaliser une analyse pondérée au moyen de ces logiciels réduit quelque peu la distorsion qui affecte les estimations ponctuelles, mais, même alors, les erreurs d'échantillonnage des estimations ponctuelles sont souvent très sous-estimées car, généralement, les procédures d'estimation de la variance ne prennent pas en considération d'autres caractéristiques de la conception, comme la stratification et la mise en grappes, de sorte que les déductions tirées de telles analyses seraient trompeuses. Par exemple, les différences entre groupes pourraient être considérées à tort comme significatives, ou bien des hypothèses pourraient être rejetées de façon injustifiée. Des déductions erronées des analyses des données pourraient, par exemple, avoir d'importantes incidences sur l'allocation des ressources et la formulation des politiques aux échelons national et régional.

91. Nous utiliserons maintenant un exemple tiré de Brogan (2004) pour illustrer le fait que l'utilisation de logiciels statistiques standard peut produire des estimations ponctuelles faussées, des erreurs types et des intervalles de confiance inappropriés et des tests de signification trompeurs. Cet exemple est fondé sur une série de données provenant d'une enquête par sondage sur la couverture des programmes de vaccination contre le tétanos toxoïde réalisée au Burundi en 1989. L'un des objectifs de l'enquête était de comparer la séropositivité [c'est-à-dire la présence de l'antitoxine du tétanos avec un titre d'au moins 0,01 unité internationale/millilitre (UI/ml)] et les programmes passés de vaccination. Pour plus amples informations sur la méthodologie et les résultats de cette enquête, voir Brogan (2004) et les ouvrages qui y sont cités. Le tableau 7.16 ci-après présente des estimations du pourcentage de femmes ayant testé positif et l'erreur type et l'intervalle de confiance connexes.

92. Il y a lieu de noter que les estimations ponctuelles sont identiques pour tous les programmes qui utilisent des pondérations, mais qu'il y a clairement des différences entre les estimations pondérées et non pondérées. De plus, les erreurs types produites par le logiciel approprié sont près de deux fois plus élevées que celles qui sont produites par le logiciel standard reposant sur l'hypothèse d'un échantillonnage aléatoire simple. Autrement dit, les logiciels standard sous-estiment gravement les variances des estimations, ce qui risque d'avoir des incidences importantes sur la formulation des politiques. Par exemple, s'il était envisagé de mettre sur pied une intervention si l'incidence de la séropositivité était égale ou inférieure à 65 %, une telle intervention serait entreprise à la suite d'une analyse faite au moyen de logiciels spéciaux mais pas sur la base d'une analyse utilisant des logiciels standard. Le tableau 7.16 fait apparaître que les logiciels qui estiment comme il convient les variances des estimations produisent approximativement les mêmes résultats. Dans la section suivante, nous donnerons un bref aperçu de certains logiciels statistiques disponibles dans le commerce qui peuvent être utilisés pour l'analyse des données provenant d'enquêtes sur les ménages.

7.9. Utilisation de logiciels pour l'estimation des erreurs d'échantillonnage

93. Les méthodes d'estimation des erreurs d'échantillonnage indiquées ci-dessus sont utilisées depuis longtemps dans les pays développés et sont employées principalement au moyen d'algorithmes informatiques individualisés mis au point par les offices nationaux de statistique, les instituts de recherche et des organismes d'enquête privés. Les progrès récents de l'informatique ont débouché sur la mise au point de plusieurs logiciels qui permettent d'appliquer ces techniques. Beaucoup d'entre eux peuvent aujourd'hui être utilisés sur des ordinateurs personnels. Ils n'utilisent que l'une des approches générales de l'estimation de la variance dont il est question à la section 7.7. La plupart de ces logiciels donnent les estimations les plus largement utilisées, comme les moyennes, les proportions, les ratios et les coefficients de régression linéaire. Certains logiciels donnent également des approximations pour une large gamme d'estimateurs, comme les coefficients de régression logistique.

94. Nous présenterons dans cette section un bref aperçu de certains des logiciels disponibles dans le commerce qui permettent d'estimer les erreurs d'échantillonnage qui caractérisent les données d'enquêtes sur les ménages. Loin de nous l'idée de vouloir donner une liste exhaustive de tous les programmes et logiciels disponibles, et nous avons simplement mentionné quelques-uns des logiciels statistiques qui peuvent actuellement être utilisés sur ordinateurs personnels par des analystes non spécialisés. Chaque logiciel est brièvement passé en revue, avec une indication des conceptions auxquelles il peut être appliqué et des méthodes d'estimation de la variance. Les avantages et les inconvénients de chaque logiciel sont également mentionnés. Nous n'avons pas essayé d'exposer en détail les procédures techniques ni les méthodes de calcul qui sont à la base de ces logiciels. Ces informations peuvent être obtenues en consultant les sites web des concepteurs et certaines des références citées à la fin du présent chapitre.

Tableau 7.16
Utilisation de plusieurs logiciels pour l'évaluation des variances des estimations provenant de l'enquête, avec la proportion de femmes ayant accouché récemment qui ont testé positif, Burundi, 1988-1989

Logiciel	Pourcentage de séropositivité	Erreur type	Intervalle de confiance de 95 %
SAS 8.2 MEANS sans pondérations	74,9	2,1	(70,8, 79,0)
SAS 8.2 MEANS avec pondérations	67,2	2,3	(62,7, 71,7)
SAS 8.2 SURVEYMEANS	67,2	4,3	(58,8, 75,6)
SUDAAN 8.0	67,2	4,3	(58,8, 75,6)
STATA 7.0	67,2	4,3	(58,8, 75,6)
EPI INFO 6.04d	67,2	4,3	(58,8, 75,6)
WESVAR 4.1	67,2	4,3	(58,8, 75,6)

95. Les six logiciels examinés sont les suivants : CENVAR, EPI INFO, PC CARP, STATA, SUDAAN et WESVAR. Les logiciels SUDAAN (Shah, Barnwell et Bieler, 1996), STATA (StataCorp, 1996), PC CARP (Fuller *et al.,* 1989) et CENVAR utilisent tous la méthode de linéarisation pour estimer les erreurs d'échantillonnage de statistiques non linéaires. Le programme WESVAR n'utilise que des méthodes de réplication. Les versions récentes du logiciel SUDAAN utilisent également les techniques BRR et « jackknife ». En outre, les logiciels SAS et SPSS (qui ne sont pas examinés ici) comportent de nouveaux modules pour l'analyse des données d'enquête. Les programmes

à réplication offrent nombre des méthodes de base, sauf la méthode « bootstrap ». Nous ne donnerons qu'une brève comparaison des attributs généraux de ces logiciels car une comparaison détaillée exigerait des comparaisons plus approfondies d'enquêtes par sondage de différentes envergures et beaucoup plus de statistiques, ce qui sortirait du champ limité de la présente étude.

96. Des liens avec les concepteurs des divers logiciels examinés et bien d'autres se trouvent à l'adresse suivante : *www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.*

97. Brogan (2004) donne une comparaison détaillée de plusieurs logiciels statistiques, y compris ceux qui sont examinés ici, sur la base des données provenant de l'enquête sur les ménages réalisée au Burundi.

98. L'on trouvera ci-après un bref aperçu de ces divers logiciels. Le lecteur intéressé trouvera des informations plus détaillées dans des manuels ou en se référant aux sites web indiqués ci-après.

CENVAR
Bureau of the Census des États-Unis, contact : International Programs Center
United States Bureau of the Census
Washington, D.C. 20233-8860
Courriel : IMPS@census.gov
Adresse électronique : *www.census.gov/ipc/www/imps*

99. CENVAR est l'un des composants d'un logiciel statistique conçu par le Bureau of the Census des États-Unis pour le traitement, la gestion et l'analyse de données d'enquêtes par sondage complexes; ce système de traitement micro-informatique intégré est appelé Integrated Microcomputer Processing System (IMPS); il est applicable à la plupart des conceptions, comme l'échantillonnage aléatoire simple, l'échantillonnage aléatoire stratifié et l'échantillonnage en grappes à plusieurs étapes. Le système CENVAR utilise une méthode d'approximation par linéarisation pour estimer la variance.

100. Les estimations produites par le système CENVAR sont notamment des moyennes, des proportions et des totaux pour l'ensemble de l'échantillon ainsi que pour des sous-groupes spécifiés, sous forme de tableaux. Indépendamment de l'erreur d'échantillonnage, le système fournit également les limites de l'intervalle de confiance de 95 %, les coefficients de variation, les effets de conception et les tailles des échantillons non pondérés.

EPI INFO
Centers for Disease Control and Prevention des États-Unis
Epidemiology Program Office, Mailstop C08
Centers for Disease Control and Prevention
Atlanta, GA 30333
Courriel : EpiInfo@cdc1.cdc.gov
Adresse électronique : *http://www.cdc.gov/epiinfo/*

101. EPI INFO est un logiciel statistique conçu par les Centers for Disease Control and Prevention des États-Unis pour le traitement, la gestion et l'analyse des données épidémiologiques, y compris des données provenant d'enquêtes complexes (élément CSAMPLE). La documentation pertinente est disponible en ligne et peut être imprimée chapitre par chapitre. Ce système est conçu spécifiquement pour l'échantillonnage en grappes stratifié à plusieurs étapes suivant le modèle d'échantillonnage fondé sur la grappe ultime.

102. EPI INFO produit des estimations de l'erreur d'échantillonnage des moyennes et des proportions pour l'ensemble de l'échantillon ainsi que pour les sous-catégories spécifiées. La sortie d'imprimante donne notamment les fréquences non pondérées, les proportions ou moyennes pondérées, les erreurs types, les limites de l'intervalle de confiance de 95 % et les effets de conception.

PC CARP
Iowa State University
Statistical Laboratory
219 Snedecor Hall
Ames, IA 50011
Adresse électronique : *http://cssm.iastate.edu/software/pccarp.html*

103. PC CARP est un programme statistique mis au point par l'Université de l'État de l'Iowa pour estimer les erreurs types des moyennes, proportions, quantiles, ratios, différences de ratios et analyse de tableaux de contingence à deux entrées. Le programme est conçu de manière à pouvoir analyser des échantillons en grappes stratifiés à plusieurs étapes. PC CARP utilise pour estimer la variance la méthode de la linéarisation.

STATA
Stata Corporation
702 University Drive East
College Station, TX 77840
Courriel : stata@stata.com
Adresse électronique : *http://www.stata.com*

104. STATA est un logiciel d'analyse statistique conçu pour estimer les erreurs d'échantillonnage de moyennes, totaux, ratios, proportions, régressions linéaires, régressions logistiques et procédures d'analyse probit. Le système permet également d'estimer des combinaisons linéaires de paramètres et de tests d'hypothèse ainsi que d'estimer des quantiles, d'analyser des tableaux de contingence, de compenser les données manquantes et de réaliser d'autres analyses. Le système STATA utilise la méthode de linéarisation pour estimer la variance.

SUDAAN
Statistical Software Center
Research Triangle Institute
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194
Courriel : SUDAAN@rti.org
Adresse électronique : *http://www.rti.org/patents/sudaan.html*

105. SUDAAN est un logiciel statistique d'analyse de données corrélées et de données provenant de conceptions complexes. Ce système peut être appliqué à des conceptions extrêmement diverses, y compris l'échantillonnage aléatoire simple et l'échantillonnage stratifié à phases multiples. Il permet d'estimer différentes statistiques et les erreurs d'échantillonnage complexes, dont moyennes, proportions, ratios, quantiles, tabulations croisées et odds ratios, modèles de régressions linéaires, logistiques et proportionnelles et analyse de tableaux de contingence. Le programme estime la variance au moyen de l'approche de linéarisation.

WESVAR
Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850-3129

Courriel : WESVAR@westat.com
Adresse électronique : *http://www.westat.com/wesvar/*

106. WESVAR est un logiciel statistique conçu par Westat, Inc. pour l'analyse de données d'enquêtes complexes, y compris l'analyse des tableaux de contingence, la régression et la régression logistique. Il peut être utilisé pour la plupart des conceptions mais il est spécifiquement conçu pour un échantillonnage en grappes stratifié à phases multiples fondé sur le modèle d'échantillonnage de la grappe ultime.

107. WESVAR utilise pour estimer la variance des méthodes de réplication, dont la méthode « jackknife », demi-échantillon équilibré et version Fay modifiée de la méthode du demi-échantillon équilibré. Il faut cependant créer une nouvelle version de la série de données sous forme spéciale WESVAR et spécifier les pondérations des réplicats.

7.10. Comparaison générale des systèmes de logiciels

108. Les systèmes de logiciels examinés ci-dessus ont beaucoup de caractéristiques communes. Tous les programmes exigent une spécification des pondérations, des données et des unités d'échantillonnage pour chaque élément. Ils ne sont pas universellement applicables à toutes les conceptions concevables. Par exemple, dans la plupart des conceptions d'échantillonnage stratifié à phases multiples, les unités primaires d'échantillonnage sont sélectionnées sur la base de probabilités proportionnelles à la taille sans remplacement. Seul un programme de la liste, SUDAAN, peut être utilisé pour ce type particulier d'enquête. Toutefois, tous les programmes peuvent être utilisés avec une telle conception aussi longtemps qu'est employé un modèle de sélection fondé sur la grappe ultime (voir la section 7.7). En outre, le logiciel SUDAAN comporte également des caractéristiques conçues de manière à estimer la variance pour des conceptions fondées sur une sélection sans remplacement d'unités d'échantillonnage primaires. Le système STATA est le seul qui comporte des fonctions d'estimation tenant compte des méthodes de stratification et de sélection à phase multiples.

109. Tous les logiciels donnent des estimations des variances d'échantillonnage et des statistiques connexes (effets de conception, corrélation intra-classe, etc.) pour les moyennes, totaux et proportions de l'ensemble de l'échantillon, pour des sous-groupes de l'ensemble de l'échantillon et pour les différences entre sous-classes. La plupart d'entre eux donnent des estimations des variances d'échantillonnage pour les statistiques obtenues par régression et par régression logistique. Tous donnent des estimations des statistiques de contrôle sur la base des variances d'échantillonnage produites.

110. CENVAR, EPI INFO, PC CARP et WESVAR sont disponibles gratuitement moyennant un droit modique. Les utilisateurs intéressés peuvent obtenir un complément d'informations sur l'acquisition des logiciels et de la documentation connexe en utilisant les adresses électroniques et autres informations fournies ci-dessus.

7.11. Conclusions

111. Le présent chapitre a donné un bref aperçu des procédures de calcul des erreurs d'échantillonnage affectant les estimations provenant de conceptions standard ou de conceptions plus complexes utilisées pour les enquêtes sur les ménages. Le calcul des erreurs d'échantillonnage est un aspect capital de l'analyse des résultats des enquêtes. Idéalement, les erreurs d'échantillonnage devraient être calculées pour toutes les caractéristiques des données. Dans la pratique, cependant, il est sélectionné

une série de caractéristiques clés pour calculer les erreurs d'échantillonnage dans chaque domaine. Les caractéristiques sélectionnées doivent être celles qui sont considérées comme les plus importantes dans le contexte de l'enquête, mais elles doivent également comprendre une sélection représentative de facteurs présentant certaines propriétés statistiques, à savoir les éléments dont on pense qu'ils sont très regroupés (par exemple des variables indiquant l'origine ethnique ou l'accès aux services), ou peu groupés (comme la situation conjugale). En outre, le choix devra être guidé par d'autres aspects comme les caractéristiques propres à une proportion élevée ou peu élevée de la population ou des domaines présentant un intérêt particulier.

112. Le présent chapitre a également préconisé l'utilisation de logiciels spécialisés pour estimer les erreurs d'échantillonnage qui caractérisent les données d'enquête. Il a été donné des exemples de situations dans lesquelles l'utilisation de logiciels statistiques standard introduit de graves erreurs dans l'estimation des erreurs d'échantillonnage. En général, l'utilisation de tels logiciels pour analyser des données d'enquêtes sur les ménages conduira à sous-estimer la variabilité réelle des estimations. Ces estimations plus réduites de l'erreur type peuvent conduire à tirer des conclusions trompeuses des résultats de l'enquête, par exemple en conduisant à conclure à d'importantes différences entre les moyennes de deux groupes ou à rejeter à tort une hypothèse.

113. L'on a également donné une liste de certains des logiciels statistiques disponibles dans le commerce, avec une indication de leur source et de leurs applications. Dans les pays en développement, le manque de connaissances ou d'expérience en matière d'estimations des erreurs d'échantillonnage est l'un des obstacles qui empêchent d'analyser en profondeur les données recueillies. Beaucoup d'analystes ne savent pas qu'il faut utiliser des logiciels spécialisés et, s'ils le savent, préfèrent ne pas apprendre à utiliser un nouveau système.

114. Il importe de souligner que ce chapitre n'est qu'une introduction et que l'estimation de la variance de données d'enquêtes complexes est un domaine très vaste et qui ne cesse de s'étendre. Le lecteur est encouragé à se référer à certains ouvrages cités à la fin du chapitre pour obtenir des informations plus détaillées et plus systématiques. Pour une analyse plus approfondie de ces logiciels et des autres systèmes disponibles, y compris des codes d'ordinateurs et des produits de certains des logiciels disponibles, voir Brogan (2004) et les ouvrages qui y sont cités.

115. Enfin, force est de reconnaître que, par suite des progrès rapides de la technologie, beaucoup de logiciels se trouvent rapidement dépassés ou comportent de nouvelles fonctions autres que celles qui sont indiquées ci-dessus. En fait, il se peut que certaines des spécifications susmentionnées soient déjà obsolètes lorsque ce guide sera publié. Il importe par conséquent de ne pas perdre de vue que les informations les plus exactes concernant les logiciels à utiliser doivent être obtenues en se référant aux manuels ou aux sites web pertinents.

Références et autres lectures

An, A. et D. Watts (2001). *New SAS procedures for analysis of sample survey data.* SUGI Paper, n° 23, Cary, North Carolina, SAS Institute, Inc. Disponible à l'adresse *http://support.sas.com/rnd/app/papers/survey.pdf.*

Binder D. A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review,* vol. 51, p. 279-92.

Brick J. M. *et al.* (1996). *A User's Guide to WesVarPC.* Rockville, Maryland, Westat, Inc.

Brogan, Donna (2005). *Estimation de l'erreur d'échantillonnage pour les données d'enquête.* Enquêtes sur les ménages dans les pays en développement et les pays en transition. Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Burt, V. L. et S. B. Cohen (1984). A comparison of alternative variance estimation strategies for complex survey data. *Proceedings of the American Statistical Association Survey Research Methods Section.*

Carlson, B. L., A. E. Johnson et S. B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics* 9, n° 4, p. 795-814.

Cochran, W. (1977). *Sampling Techniques,* troisième édition. New York, Wiley & Sons.

Dippo, C. S., R. E. Fay et D. H. Morganstein (1984). Computing variances from complex samples with replicate weights. *Proceedings of the American Statistical Association Survey Research Methods Section.*

Fuller, Wayne *et al.* (1989). PC CARP: USERS MANUAL. Ames, Iowa, Statistics Laboratory, Iowa State University. Disponible à l'adresse *http://cssm.lastate.edu/software*.

Hansen, M. H., W. N. Hurwitz et W. G. Madow (1953). *Sample Survey Methods and Theory,* vol. I, *Methods and Applications*. New York, Wiley & Sons, Sect. 10.16.

Hansen M. H., W. G. Madow et B. J. Tepping (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association,* vol. 78, n° 384, p. 776-793.

Kish, Leslie (1965). *Survey Sampling*. New York, John Wiley and Sons.

_____ (1995). *Leslie Kish: Selected Papers,* Steven Heeringa and Graham Kalton, eds. Hoboken, New Jersey, John Wiley & Sons, Inc.

Kish, L. et M. R. Frankel (1974). Inference from complex samples. *Journal of the Royal Statistical Society: services B,* vol. 36, p. 1-37.

Landis J. R. *et al.* (1982). A statistical methodology for analyzing data from a complex survey: the First National Health and Nutrition Examination Survey. *Vital and Health Statistics,* vol. 2, n° 92. Washington, Department of Health, Education and Welfare.

Lehtonen, R. et E. J. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York, Wiley & Sons.

Lepkowski J. M., J. A. Bromberg et J. R. Landis (1981). *A program for the analysis of multivariate categorical data from Complex Sample Surveys.* Proceedings of the American Statistical Association Statistical Computing Section.

Levy, Paul S. et Stanley Lemeshow (1999). *Sampling of Populations: Methods and Applications*. Troisième édition. New York, John Wiley & Sons.

Lohr, Sharon (1999). *Sampling: Design and Analysis.* Pacific Grove, California, Duxbury Press.

Organisation des Nations Unies (1993). *Sampling errors in household surveys.* UNFPA/UN/INT-92-P80-15E. New York, Division de statistique, Département de l'information économique et sociale et de l'analyse des politiques, et Programme de mise en place de dispositifs nationaux d'enquête sur les ménages.

Potthoff, R. F., M. A. Woodbury et K. G. Manton (1992). « Equivalent sample size » and « equivalent degrees of freedom » refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association,* vol. 87, p. 383-396.

Rust K. (1985). « Variance Estimation for Complex Estimators in Sample Surveys ». *Journal of Official Statistics* 1(4), 381-397.

Rust K. F. et J. N. K. Rao (1996). Variance estimation for complex surveys using replication techniques », *Statistical Methods in Medical Research,* vol. 5, p. 283-310.

Shah, Babhai V. (1998). Linearization methods of variance estimation. *Encyclopedia of Biostatistics,* vol. 3, Peter Armitage and Theodore Colton, eds. New York, John Wiley and Sons, p. 2276-2279.

Shah B. V., B. G. Barnwell et G. S. Bieler (1996). *SUDAAN User's Manual: Release 7.0.* Research Triangle Park, North Carolina, Research Triangle Institute.

Tepping B. J. (1968). Variance estimation in complex surveys. *Proceedings of the American Statistical Association Social Statistics Section,* p. 11-18.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* New York, Springer-Verlag.

Woodruff R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association,* vol. 66, n° 334, p. 411-414.

Chapitre 8

Erreurs autres que les erreurs d'échantillonnage dans les enquêtes sur les ménages

8.1. Introduction

1. Les erreurs d'échantillonnage, de même que les *erreurs autres que d'échantillonnage,* doivent être maîtrisées et ramenées à un niveau tel que leur présence n'ôte pas toute utilité aux résultats finals de l'enquête. Les chapitres précédents, consacrés à la conception de l'échantillon et aux méthodes d'estimation, ont porté principalement sur l'erreur d'échantillonnage et un peu moins sur les autres sources d'écart, comme la non-réponse et la non-couverture, qui constituent la catégorie d'erreurs appelées collectivement *erreurs autres que d'échantillonnage*. Les erreurs autres que d'échantillonnage sont particulièrement préjudiciables lorsqu'elles ne sont pas aléatoires en raison de la distorsion qu'elles introduisent dans les estimations.

2. Toutes les données d'enquête sont sujettes à des erreurs provenant de différentes sources. La distinction fondamentale de caractère général est à établir entre les erreurs qui surgissent lors des processus de mesure et les erreurs d'échantillonnage, c'est-à-dire les erreurs qui affectent l'estimation des paramètres de la population tirée de la mesure d'un échantillon de celle-ci.

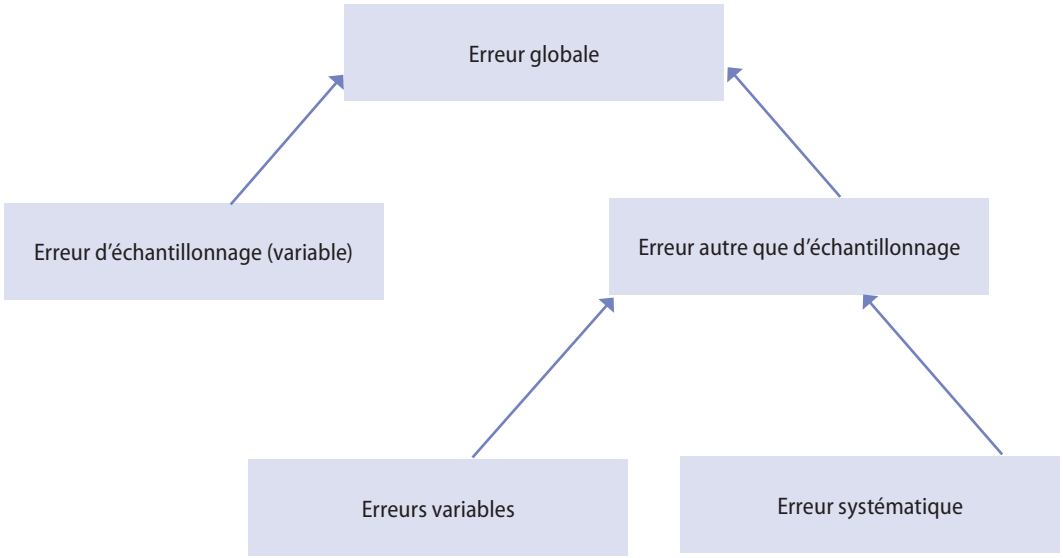
3. Dans les chapitres précédents, il a été pris pour hypothèse que chaque unité Y_i d'une population était associée à une valeur y_i réputée être la valeur réelle de l'unité pour la caractéristique y. Il a également été pris pour hypothèse que, alors que y_i faisait partie de l'échantillon, la valeur de y signalée ou observée était de y_i. Cela est vrai dans certaines situations, mais pas dans toutes. Par exemple, dans les pays où il existe un système viable et complet d'enregistrement des statistiques de l'état civil fondé sur les certificats de naissance, les valeurs « réelles » peuvent être aisément obtenues lorsque y_i dénote l'âge. Dans d'autres cas, cependant, lorsqu'il s'agit par exemple de porter un jugement qualitatif sur son propre état de santé, il peut être un peu plus difficile d'obtenir des valeurs réelles ou même de les définir. Il se peut par exemple qu'un malade se considère comme étant en bonne santé, selon les circonstances.

4. Dans la pratique, l'on ne peut pas tenir pour acquis que la valeur signalée observée pour l'unité Y_i est toujours égale à y_i, quelle que soit la source de l'information ou les circonstances dans lesquelles l'information est obtenue. L'expérience offre de nombreux exemples qui montrent que des erreurs de mesure ou d'observation ainsi que des erreurs dues à des réponses erronées, à la non-réponse et à d'autres causes peuvent surgir lors de l'enquête.

5. Indépendamment des erreurs de réponse, il peut y avoir une erreur de couverture, de traitement, etc. La qualité de l'estimateur d'un paramètre de la population est fonction de l'erreur totale,

c'est-à-dire aussi bien des erreurs d'échantillonnage que des erreurs autres que d'échantillonnage. Comme on l'a déjà dit, les erreurs d'échantillonnage sont imputables exclusivement à la sélection d'un échantillon probabiliste plutôt qu'à la réalisation d'une énumération complète. Les erreurs autres que d'échantillonnage, en revanche, sont dues principalement aux procédures de collecte et de traitement des données. La figure 8.1 illustre la corrélation entre les erreurs d'échantillonnage et les erreurs autres que d'échantillonnage en tant que composantes de l'erreur globale.

Figure 8.1
Corrélation entre les erreurs d'échantillonnage et les erreurs autres que d'échantillonnage en tant qu'éléments de l'erreur globale



6. Les *erreurs autres que d'échantillonnage,* par conséquent, sont surtout le résultat de définitions et de concepts non valables, de cadres d'échantillonnage inexacts, de questionnaires mal conçus, de méthodes défectueuses de collecte, de tabulation et de codage des données et d'une couverture incomplète des unités d'échantillonnage. Ces erreurs sont imprévisibles et il est difficile de les maîtriser. À la différence des *erreurs d'échantillonnage,* ce type d'erreur peut augmenter parallèlement à la taille de l'échantillon. Si elles ne sont pas maîtrisées comme il convient, les *erreurs autres que d'échantillonnage* peuvent avoir un effet plus néfaste que les erreurs d'échantillonnage dans le cas d'enquêtes de grande envergure sur les ménages.

8.2. Distorsion et erreur variable

7. Comme le montre le tableau 8.1, les erreurs d'échantillonnage peuvent être décomposées en erreurs variables ou en distorsion. Les erreurs variables sont dues principalement aux erreurs d'échantillonnage, mais des erreurs autres que d'échantillonnage, essentiellement celles qui sont connues lors des opérations de traitement, comme le codage et l'entrée des données, peuvent également contribuer aux erreurs variables. En revanche, la distorsion est imputable principalement à des erreurs autres que d'échantillonnage dues à des éléments comme des définitions non valables, des procédures erronées de mesure, des réponses erronées, la non-réponse, la couverture incomplète de la population cible,

etc. Certains types de distorsions peuvent également être dues aux erreurs d'échantillonnage : elles peuvent provenir par exemple d'un calcul des variances d'échantillonnage effectué sur la base d'un estimateur de la variance qui ne reflète pas comme il convient la conception de l'échantillon et qui entraîne par conséquent une surestimation ou une sous-estimation des erreurs d'échantillonnage.

8. L'on entend généralement par distorsion les erreurs systématiques qui affectent les enquêtes réalisées sur la base d'une conception spécifiée avec la même erreur constante. Comme indiqué ci-dessus, les erreurs d'échantillonnage constituent habituellement la source de la plupart des erreurs variables, tandis que les distorsions découlent essentiellement d'*erreurs autres que d'échantillonnage*. Ainsi, les distorsions proviennent des défaillances de la conception fondamentale et des procédures d'enquête tandis que les erreurs variables sont imputables au fait que les conceptions et procédures d'enquête n'ont pas été systématiquement appliquées.

Tableau 8.1
Classification des erreurs d'enquête

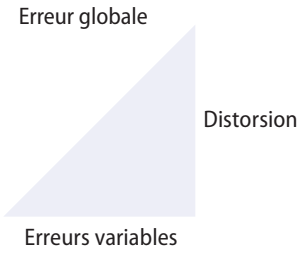
Erreurs variables	Erreur d'échantillonnage
	Erreur autre que d'échantillonnage
Distorsion	Erreur d'échantillonnage
	Erreur autre que d'échantillonnage

9. Le terme statistique qui désigne l'erreur globale est le *carré moyen d'erreur,* qui est égal à la variance plus le carré de la distorsion (voir la figure 8.2). Si, en tant qu'hypothèse d'école, la distorsion était égale à zéro, le carré moyen serait donc simplement la variance de l'estimation. Dans les enquêtes sur les ménages, toutefois, la distorsion n'est jamais égale à zéro. Comme indiqué ci-dessus, toutefois, la mesure de la distorsion totale dans les enquêtes est virtuellement impossible, notamment parce qu'il faut, pour la calculer, savoir quelle est la valeur réelle de la population, laquelle est généralement inconnue. Les sources de distorsion sont si nombreuses et elles sont si complexes qu'il est rare que l'on essaie de les estimer globalement.

10. Le triangle de la figure 8.2 ci-dessous illustre l'erreur globale et ses composantes. La hauteur du triangle représente les distorsions et la base l'erreur variable. Le fait que l'hypoténuse est la mesure de l'erreur globale reflète le théorème selon lequel la racine du carré moyen d'erreur (c'est-à-dire l'erreur globale) est égale à la racine carrée du produit de la variance d'échantillonnage plus la distorsion au carré. Par conséquent :

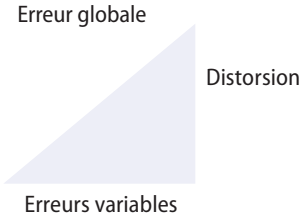
$$\text{Racine du carré moyen d'erreur} = \sqrt{VE^2 + distorsion} \qquad (8.1)$$

Figure 8.2
Erreur globale et ses composantes



11. Si l'on réduit l'erreur variable ou distorsion, l'erreur globale se trouve réduite en conséquence. La figure 8.3 illustre la situation dans laquelle aussi bien l'erreur variable que la distorsion sont considérablement réduites. L'erreur globale l'est donc aussi, comme le montre la longueur de l'hypoténuse par rapport à la longueur de celle qui est illustrée à la figure 8.2.

Figure 8.3
Réduction de l'erreur globale d'enquête



12. Le but que visent une bonne conception de l'échantillon et une bonne stratégie de réalisation d'une enquête est de réduire aussi bien l'erreur variable que la distorsion pour obtenir des résultats relativement exacts.

13. D'une manière générale, les échantillons de grande taille et les échantillons qui sont par ailleurs judicieusement conçus donnent un degré de précision élevé, tandis que des résultats exacts ne peuvent être obtenus que si aussi bien l'erreur variable que la distorsion sont réduites au minimum. Cela signifie qu'une conception précise peut néanmoins être très inexacte si la distorsion est marquée. Il importe de ne pas perdre de vue, dans ce contexte, que les estimations des erreurs standard qui sont fréquemment données dans les rapports sur les enquêtes sur les ménages sous-estiment l'erreur globale car elles ne tiennent pas compte de l'impact de la distorsion.

14. Dans la pratique, les erreurs autres que d'échantillonnage peuvent être décomposées en un élément variable et en des erreurs systématiques (Biemer et Lyberg, 2003). D'une manière générale, les erreurs systématiques ne se compensent pas et aident par conséquent à converger (essentiellement dans la même direction), tandis que les erreurs variables se compensent et ont tendance à être contraires, c'est-à-dire à s'annuler.

8.2.1. Élément variable

15. L'élément variable d'une erreur est dû à des facteurs aléatoires qui affectent différents échantillons et la répétition de l'enquête. Dans le cas du processus de mesure, nous pouvons imaginer que

l'ensemble des procédures, lors de la sélection des enquêteurs à la collecte et au traitement des données peuvent être répétées au moyen des mêmes méthodes spécifiées, dans les mêmes conditions données de façon indépendante sans qu'une répétition n'affecte une autre. Les résultats des répétitions sont affectés par des facteurs aléatoires ainsi que par des facteurs systématiques qui tiennent aux conditions dans lesquelles les répétitions sont entreprises et qui affectent les résultats de la répétition de la même façon.

16. Lorsque les erreurs variables sont dues uniquement à des erreurs d'échantillonnage, le carré de l'erreur variable est égal à la variance d'échantillonnage. L'écart entre la valeur moyenne tirée de l'enquête et la valeur réelle de la population est la distorsion. Aussi bien les erreurs variables que les distorsions peuvent provenir d'opérations d'échantillonnage ou autres que d'échantillonnage. L'erreur variable mesurera l'écart entre l'estimateur et sa valeur escomptée et comprendra aussi bien la variance d'échantillonnage que la variance autre que d'échantillonnage. L'écart entre la valeur escomptée de l'estimateur et sa valeur réelle est la distorsion totale et comprend à la fois la distorsion d'échantillonnage et la distorsion autre que d'échantillonnage.

17. Les erreurs variables peuvent être évaluées sur la base de comparaisons judicieusement conçues entre les répétitions (réplications) des opérations d'enquête dans les mêmes conditions. Pour réduire les erreurs variables, il ne suffit pas d'accroître la taille de l'échantillon et d'utiliser un plus grand nombre d'enquêteurs. D'un autre côté, la distorsion ne peut être réduite qu'en améliorant les procédures d'enquête, par exemple en introduisant des mesures de contrôle de la qualité aux différentes phases de l'enquête.

8.2.2. Erreur systématique (distorsion)

18. Il se produit une erreur systématique lorsque, par exemple, il existe une tendance à sous-estimer ou à surestimer systématiquement les données. Par exemple, dans certains pays où il n'existe pas de certificats d'enregistrement des naissances, les hommes ont tendance à se déclarer comme étant plus âgés qu'ils ne le sont réellement. Cette pratique entraînerait manifestement une distorsion systématique, c'est-à-dire une surestimation de l'âge moyen de la population de sexe masculin.

8.2.3. Distorsion d'échantillonnage

19. La distorsion d'échantillonnage peut être due à une sélection inadéquate ou défectueuse de l'échantillon probabiliste spécifié ou à des méthodes d'estimation défectueuses. Dans le premier cas, il peut s'agir de défauts des cadres d'échantillonnage, de procédures de sélection erronées et d'une énumération partielle ou incomplète des unités sélectionnées. Les chapitres 3 et 4 du présent guide contiennent une discussion détaillée des nombreuses circonstances dans lesquelles une application inadéquate d'une conception d'échantillonnage – même presque parfaite – peut entraîner une distorsion.

8.2.4. Comparaison de la distorsion et de l'erreur variable

20. D'une manière générale, les distorsions sont difficiles à mesurer, et c'est pourquoi nous insistons sur la nécessité de tout faire pour les minimiser. Il n'est possible de les évaluer qu'en comparant les résultats de l'enquête et des sources de données externes et fiables. D'un autre côté, l'erreur variable peut être évaluée en comparant des sous-groupes de l'échantillon ou en répétant l'enquête dans de meilleures conditions. La distorsion peut être réduite en améliorant les procédures d'enquête.

21. Selon Verma (1991), les erreurs provenant des mêmes sources, dont la couverture, la non-réponse et la sélection de l'échantillon, revêtent principalement la forme d'une distorsion. D'un autre côté, les erreurs de codage d'entrée des données peuvent apparaître essentiellement comme une erreur variable.

22. Les erreurs systématiques comme les erreurs variables affectent l'exactitude et la fiabilité globales des résultats, mais la distorsion touche davantage des estimations comme les moyennes, proportions et totaux de la population. Ces estimations linéaires sont la somme des observations concernant l'échantillon. Comme on l'a déjà vu, aussi bien les erreurs variables autres que l'échantillonnage que les erreurs d'échantillonnage peuvent être réduites en augmentant la taille de l'échantillon. Dans le cas d'estimations non linéaires comme les coefficients de corrélation, les erreurs types et les estimations issues de régressions, aussi bien les erreurs variables que les erreurs systématiques peuvent entraîner de graves distorsions (Biemer et Lyberg, 2003). Comme, dans le cas de beaucoup d'enquêtes sur les ménages, le principal objectif est d'obtenir des mesures descriptives de la population comme des moyennes, des totaux et des proportions, il faut s'efforcer surtout de réduire les erreurs systématiques.

23. En résumé, la distorsion découle des défaillances de la conception fondamentale et des procédures d'enquête. Elle est plus difficile à mesurer que l'erreur variable et ne peut être évaluée que sur la base d'une comparaison avec des sources d'information plus fiables indépendantes de l'enquête au moyen d'informations obtenues grâce à des procédures améliorées.

8.3. Sources d'erreurs autres que d'échantillonnage

24. Les différentes causes d'erreurs autres que d'échantillonnage, et elles sont nombreuses, existent d'emblée, au stade de la planification et de la conception de l'enquête, jusqu'à la dernière étape, lorsque les données sont traitées et analysées.

25. Un programme d'enquêtes sur les ménages peut être considéré comme une série de règles rigoureuses qui spécifient différentes opérations. Ces règles, par exemple, décrivent la population cible à étudier, spécifient les concepts thématiques et définitions à utiliser dans le questionnaire et exposent les méthodes à suivre pour rassembler les données et procéder aux mesures. Si les opérations d'enquête sont réalisées conformément aux règles fixées, il est théoriquement possible d'obtenir une *valeur réelle* de la caractéristique à l'étude. Cependant, du fait des erreurs autres que d'échantillonnage, il s'agit là d'un idéal irréalisable.

26. D'une manière générale, les erreurs autres que d'échantillonnage peuvent être causées par un ou plusieurs des facteurs suivants :

a) Inadéquation et/ou manque de cohérence de la spécification des données en ce qui concerne les objectifs de l'enquête;

b) Double décompte ou omission d'unités par suite d'une définition peu précise des limites des unités géographiques d'échantillonnage;

c) Identification incomplète ou incorrecte des unités d'échantillonnage[1] ou méthodes d'entrevues défectueuses;

1 Il y a lieu de noter que cette erreur, bien que se produisant au stade de la sélection de l'échantillon, est néanmoins un type de distorsion autre que d'échantillonnage.

d) Méthodes inappropriées d'entrevues, d'observation et de mesure dues à des questionnaires, définitions ou instructions ambigus;

e) Manque d'enquêteurs formés et expérimentés et, en particulier, absence de supervision de qualité sur le terrain;

f) Vérification insuffisante des données de base afin de corriger les erreurs évidentes;

g) Erreurs survenues lors des opérations de traitement, par exemple de codage, d'entrée, de vérification, de tabulation, etc., des données;

f) Erreurs commises lors de la présentation et de la publication des résultats de l'enquête.

Toutefois, cette liste est loin d'être exhaustive.

8.4. Éléments des erreurs autres que d'échantillonnage

27. Biemer et Lyberg (2003) ont identifié cinq éléments d'erreurs autres que d'échantillonnage, à savoir des erreurs de spécification, de cadre, de non-réponse, de mesure et de traitement. L'on peut ajouter à cette liste l'erreur d'estimation. Ces types d'erreur sont brièvement évoqués ci-dessous.

8.4.1. Erreur de spécification

28. Il y a erreur de spécification lorsque le concept qui sous-tend la question est autre que l'élément sous-jacent à évaluer. Par exemple, dans certaines cultures, une question simple, comme celle de savoir combien d'enfants une personne a eus, peut être interprétée différemment. Dans le cas des ménages composés d'une famille élargie, les enfants biologiques du déclarant peuvent ne pas être distingués des enfants des frères et sœurs qui vivent dans le même ménage. Dans le cas d'une enquête sur les handicaps, une question générale tendant à savoir si les déclarants souffrent ou non d'un handicap peut être interprétée différemment selon la gravité du handicap ou l'idée que le déclarant s'en fait. Il se peut par exemple que les personnes souffrant d'un handicap mineur ne se considèrent pas comme handicapées. Si le questionnaire ne comporte pas de filtre approprié, les réponses risquent de ne pas révéler le nombre total de personnes souffrant d'un handicap.

8.4.2. Erreur de couverture ou de cadre

29. Le plus souvent, les unités primaires d'échantillonnage comprennent des grappes d'unités géographiques comme des zones d'énumération du recensement (voir le chapitre 4 pour une étude approfondie des cadres d'échantillonnage). Or, il n'est pas inhabituel que les zones d'énumération soient mal délimitées lors de l'établissement des cartes devant servir au recensement. Il se peut ainsi que des ménages soient omis ou dénombrés deux fois lors de la deuxième phase. De telles imperfections peuvent fausser les estimations de l'enquête dans deux directions. Si les unités qui auraient dû faire partie du cadre n'en font pas partie, la probabilité de sélection des unités omises sera nulle, ce qui se traduira par une sous-estimation. D'un autre côté, si certaines unités sont dénombrées deux fois, le résultat sera une sur-couverture et par conséquent une surestimation.

30. Les erreurs liées au cadre peuvent par conséquent entraîner aussi bien une *sur-couverture* qu'une *sous-couverture*. Celle-ci est le résultat le plus commun dans le cas des enquêtes de grande envergure menées dans la plupart des pays d'Afrique.

31. Dans le cas des enquêtes sur les ménages à phases multiples, l'échantillonnage comporte plusieurs étapes, dont la sélection des unités géographiques, en une ou plusieurs phases, la sélection et l'établissement d'une liste de ménages et l'établissement d'une liste et la sélection des personnes faisant partie des ménages sélectionnés (voir chapitre 3). Une erreur de couverture peut se produire à n'importe laquelle de ces étapes.

32. Il importe de souligner à nouveau que ni l'ampleur ni l'effet des erreurs de couverture ne sont faciles à estimer car il faut pour cela disposer d'informations externes autres que celles qui proviennent de l'échantillon mais aussi, par définition, du cadre d'échantillonnage utilisé.

33. La *non-couverture* signifie, comme indiqué ci-dessus, que certaines des unités d'une population définie ne figurent pas dans le cadre d'échantillonnage (voir le chapitre 6 pour une discussion plus approfondie de l'erreur de couverture, y compris la non-couverture). Comme certaines unités ont une probabilité de sélection nulle, elles se trouvent à toutes fins utiles exclues des résultats de l'enquête.

34. Il importe de noter qu'il ne s'agit pas ici de l'exclusion délibérée et explicite de certains secteurs d'une population plus vaste. De telles exclusions délibérées peuvent être dictées par les objectifs de l'enquête ou des difficultés pratiques. Par exemple, des enquêtes sur les attitudes à l'égard du mariage peuvent exclure les personnes n'ayant pas atteint l'âge minimum légal du mariage. Les personnes qui vivent en institution sont souvent exclues en raison des difficultés pratiques qu'il y aurait à les interroger. Les régions d'un pays semées de mines terrestre peuvent être exclues d'une enquête sur les ménages pour garantir la sécurité des enquêteurs. Lorsque l'on calcule les taux de non-couverture, les membres du groupe délibérément et explicitement exclus ne doivent pas être dénombrés, que ce soit parmi la population cible ou sous la rubrique de la non-couverture. À ce propos, la définition de la population cible doit être un des éléments des conditions essentielles de l'enquête qui doivent être clairement stipulés (voir le chapitre 3 pour la question de la population cible).

35. L'expression *erreur brute de couverture* désigne la somme des valeurs absolues des taux d'*erreur de non-couverture* et de *sur-couverture*. L'*erreur nette de non-couverture* désigne l'excédent de la non-couverture par rapport à la sur-couverture et constitue par exemple leur somme algébrique. La couverture nette ne désigne la couverture brute que s'il n'y a pas de sur-couverture. La plupart des enquêtes sur les ménages menées dans les pays en développement sont affectées principalement par des erreurs de sous-couverture. La plupart des praticiens sont unanimes à reconnaître que, dans le cas de la plupart des enquêtes sociales, la sous-couverture est un problème beaucoup plus commun que celui de la sur-couverture. Des ajustements et une pondération sont beaucoup plus difficiles pour compenser une non-couverture qu'une non-réponse car les taux de couverture ne peuvent pas être tirés de l'échantillon lui-même, mais seulement de source externe.

36. Les erreurs de non-couverture peuvent être causées par l'utilisation de cadres ou d'unités d'échantillonnage défectueux, comme on l'a vu en détail au chapitre 4. Si les cadres d'échantillonnage ne sont pas mis à jour et s'il est utilisé des cadres anciens pour économiser du temps ou de l'argent, l'on risque une sérieuse distorsion. Dans une enquête sur les ménages, par exemple, si une liste de logements ancienne n'est pas mise à jour depuis qu'elle a été établie (ce qui peut être 10 ans avant l'enquête), les logements nouveaux qui sont venus s'ajouter à la zone d'énumération sélectionnée ne feront pas partie du cadre secondaire. De même, les logements abandonnés continueront de figurer dans le cadre, mais en blanc. En pareil cas, il peut y avoir simultanément une omission d'unités faisant partie de la population cible et une inclusion d'unités qui ne doivent pas en faire partie.

37. Il arrive également que certaines unités de l'échantillon ne soient pas localisées ou visitées. Ce problème peut également découler de l'utilisation de listes incomplètes. En outre, les problèmes de transport ou le mauvais temps peuvent rendre certaines unités inaccessibles pendant la période prévue pour l'enquête.

38. Comme on l'a vu au chapitre 3, l'objectif ultime d'une enquête par sondage est d'obtenir des résultats objectifs qui permettent d'opérer des déductions valables concernant la population visée à partir d'une observation des unités faisant partie de l'échantillon. Les résultats de l'enquête peuvent par conséquent être faussés si l'étendue de la non-couverture varie d'une région géographique à l'autre et d'un sous-groupe à l'autre, comme hommes et femmes, groupes d'âge, groupes ethniques et catégories socioéconomiques.

39. Les erreurs de non-couverture ne sont pas les mêmes que les erreurs de non-réponse. Ces dernières, comme on l'a dit, sont dues au fait qu'il n'a pas été obtenu d'observations pour certaines unités d'échantillonnage en raison de refus de répondre, de l'impossibilité de trouver une adresse, de l'absence temporaire de déclarants, de la perte de questionnaires, etc. L'étendue de la non-réponse est mesurée à partir des résultats de l'échantillon en comparant l'échantillon sélectionné et l'échantillon réel. Comme indiqué ci-dessus, l'étendue de la même couverture, en revanche, ne peut être évaluée qu'au moyen d'une vérification externe à l'enquête.

8.4.2.1. Erreurs de sélection de l'échantillon

40. Les erreurs de sélection de l'échantillon sont les omissions et distorsions qui caractérisent le cadre d'échantillonnage, par exemple par suite d'une application erronée des taux ou des procédures de sélection. Un autre exemple serait un remplacement inapproprié, sur le terrain, des ménages sélectionnés par d'autres ménages plus accessibles ou plus coopératifs.

8.4.2.2. Réduction de l'erreur de couverture

41. Pour réduire l'erreur de couverture, le meilleur moyen consiste à améliorer le cadre d'échantillonnage en excluant les unités qui s'y trouvent par erreur et les doubles dénombrements. Le mieux, pour cela, est de veiller à ce que des cadres existants soient dûment mis à jour (voir le chapitre 4 pour un examen détaillé de cette question). Il importe également de veiller à ce que les unités géographiques d'échantillonnage et les ménages qui s'y trouvent puissent être aisément localisés. Pour cela, il convient d'établir des cartes appropriées lors de la construction du cadre initial, habituellement lors du dernier recensement de la population et du logement.

8.4.3. Non-réponse

42. Comme on l'a noté à maintes reprises dans le présent guide, la non-réponse désigne une impossibilité d'obtenir une réponse de certaines des unités d'échantillonnage. Il est bon de se représenter la population visée comme scindée en deux strates, l'une comprenant toutes les unités d'échantillonnage pour lesquelles des réponses ont été obtenues et l'autre toutes les unités d'échantillonnage pour lesquelles il n'a pas été possible d'obtenir de réponse.

43. Le plus souvent, la non-réponse n'est pas également répartie entre les unités d'échantillonnage mais est plutôt très concentrée parmi certains sous-groupes. Du fait de ces différences, la répartition de l'échantillon réel entre les sous-groupes s'écartera de celle de l'échantillon sélectionné. Cet écart

risque d'entraîner une distorsion de la non-réponse si les variables visées par l'enquête sont également liées aux sous-groupes en question.

44. Le taux de *non-réponse* peut être estimé de façon exacte si tous les éléments admissibles qui font partie de l'échantillon sont dénombrés. Le *taux de réponse* est défini comme étant le ratio entre le nombre de questionnaires remplis pour les unités d'échantillonnage et le nombre total d'unités d'échantillonnage admissibles[2] (voir également le chapitre 6). Il est recommandé de donner des informations sur la non-réponse dans toutes les publications rendant compte des résultats de l'enquête, et cette pratique devrait être obligatoire dans toutes les enquêtes officielles. La non-réponse peut être due au fait que les personnes sélectionnées ne sont pas chez elles, refusent de participer à l'enquête ou sont, pour une raison ou pour une autre, incapables de répondre aux questions posées. La non-réponse peut être imputable aussi à la perte de questionnaires ou à l'impossibilité de mener une enquête dans certaines régions en raison du mauvais temps, de difficultés d'accès ou du manque de sécurité. Toutes les catégories de non-réponse se rattachent aux déclarants admissibles et doivent exclure ceux qui ne le sont pas, comme l'indique la note 2. Par exemple, dans le cas d'une enquête sur la fécondité, la population cible dans les zones d'énumération sélectionnées ne comprendra que des femmes en âge de procréer et exclura par conséquent les femmes d'autres groupes d'âge et tous les hommes.

45. Comme on l'a déjà vu au chapitre 6, il y a deux types de non-réponse: la *non-réponse unitaire* et la *non-réponse ponctuelle*. La *non-réponse unitaire* signifie qu'il n'a pas été obtenu d'informations d'une unité d'échantillonnage donnée, tandis que la *non-réponse ponctuelle* désigne le cas où il a été rassemblé une partie des informations requises, mais pas toutes, pour l'unité considérée. La non-réponse ponctuelle se traduit par des lacunes dans l'enregistrement des données concernant les unités déclarantes et peut être due à des refus, à une omission de l'enquêteur ou à une incapacité quelconque. Le refus d'un déclarant potentiel de participer à l'enquête peut être influencé par de nombreux facteurs, comme l'absence de motivation, le manque de temps, le caractère délicat de certaines questions, etc. Groves et Couper (1995) suggèrent un certain nombre de causes de refus, dont le contexte social de l'étude, les caractéristiques du déclarant, la conception de l'enquête (y compris le travail qu'elle représente pour le déclarant), les caractéristiques de l'enquêteur et l'interaction entre celui-ci et le déclarant. Dans le cas spécifique de la non-réponse ponctuelle, le déclarant peut considérer certaines des questions posées comme embarrassantes, délicates et/ou sans rapport avec l'objectif déclaré de l'enquête. Il se peut également que l'enquêteur saute une question ou n'enregistre pas une réponse. En outre, une réponse peut être rejetée lors du travail d'édition.

46. L'ampleur de la non-réponse unitaire (total), entre autres facteurs, reflète la réceptivité générale, la complexité, l'organisation et la gestion de l'enquête et par conséquent la complexité, la clarté et l'acceptabilité des éléments d'information demandés dans le questionnaire et la qualité du travail de l'enquêteur.

47. La non-réponse introduit dans les résultats de l'enquête une distorsion qui peut être grave dans les cas où les unités non déclarantes ne sont pas « représentatives » de celles qui ont répondu, comme cela est généralement le cas. La non-réponse accroît à la fois l'erreur d'échantillonnage, en réduisant la taille de l'échantillon, et les erreurs autres que d'échantillonnage.

2 Certaines des unités sélectionnées peuvent s'avérer être étrangères à l'enquête et par conséquent inadmissibles, comme des logements vacants, condamnés ou abandonnés.

48. Les efforts entrepris pour accroître le taux de réponse entraîneront souvent des modifications des procédures suivies, en particulier pour ce qui est du choix des opérations. C'est ainsi par exemple que, pour accroître le taux de réponse lors de l'Enquête sur la fécondité qui a été menée en Zambie en 1978, il a été recruté des institutrices comme enquêteuses pour poser les questions sur la contraception, etc., l'idée étant que si l'on avait recours à de jeunes hommes, le taux de refus de participation serait plus élevé, étant donné qu'il est inapproprié pour de jeunes hommes de poser à des femmes plus âgées qu'eux, en particulier, des questions sur des sujets sexuels comme la contraception.

49. La non-réponse ne peut pas être totalement éliminée dans la pratique, mais elle peut être minimisée par des techniques de persuasion, des visites répétées des ménages dont les membres étaient temporairement absents et d'autres méthodes. Pour de plus amples informations sur le traitement de la non-réponse ponctuelle dans les données d'enquête, voir les chapitres 6 et 9.

8.4.4. Erreur de mesure

50. Ces erreurs surgissent lorsque ce qui est observé ou mesuré s'écarte des valeurs qui sont effectivement celles des unités sélectionnées. Ces erreurs portent sur le contenu de fond de l'enquête, comme la définition des objectifs de l'enquête, leur traduction en questions utilisables et l'obtention, l'enregistrement, le codage et le traitement des réponses. Ces erreurs affectent par conséquent l'exactitude de la mesure au niveau des unités individuelles.

51. Ainsi, la création, lors de la phase initiale, de définitions et de concepts erronés ou trompeurs concernant la construction du cadre d'échantillonnage et la présentation du questionnaire affectera la complétude de la couverture et conduira différents enquêteurs à interpréter les définitions et les concepts différemment, ce qui affectera l'exactitude des données rassemblées.

52. L'erreur peut également être due au fait qu'il n'a pas été donné d'instructions appropriées au personnel de terrain. Dans certaines enquêtes, le manque de précision et de clarté des instructions conduira l'enquêteur à faire appel à son propre jugement pour mener à bien le travail de terrain. L'enquêteur lui-même peut être une source d'erreur: parfois, les informations rassemblées sur un point déterminé peuvent être inexactes pour toutes les unités, essentiellement parce que le personnel de terrain n'a pas été formé comme il convient.

53. En Afrique, les questions concernant l'âge constituent un problème de mesure fréquent par suite de l'imprécision des réponses. De telles erreurs de mesure, et bien d'autres encore, peuvent être imputables aux déclarants, à l'enquêteur ou aux deux. Parfois, l'interaction entre les déclarants et l'enquêteur peut contribuer à gonfler de telles erreurs. Les défaillances de l'appareil de la technique de mesure peuvent également entraîner des erreurs d'observation.

54. Les déclarants peuvent également introduire des erreurs :

- Lorsqu'ils ne comprennent pas la ou les questions posées.
- Lorsqu'ils donnent des réponses hâtives et incorrectes, par exemple s'ils ne comprennent pas vraiment quels sont les objectifs de l'enquête; il se peut en particulier que les déclarants ne réfléchissent pas assez à la question posée.
- En voulant « coopérer » en répondant à des questions alors même qu'ils ne savent pas quelle est la réponse correcte.

- En donnant délibérément des réponses inexactes, par exemple dans le cas d'enquêtes portant sur des questions délicates, comme leurs revenus ou des maladies suscitant l'opprobre social.
- Par suite de trous de mémoire lorsque la période de référence est longue, par exemple lorsqu'il s'agit de rassembler des informations concernant des produits non durables dans le contexte d'enquêtes sur les dépenses.

55. L'effet cumulé des diverses erreurs de différentes sources peut être considérable étant donné que ces erreurs risquent de ne pas se compenser, de sorte que leur effet net peut être une distorsion marquée.

8.4.5. Erreur de traitement

56. Les erreurs de traitement comprennent, entre autres :

- Les erreurs d'édition
- Les erreurs de codage
- Les erreurs d'entrée des données
- Les erreurs de programmation.

57. Ces erreurs surgissent à l'étape du traitement des données. Par exemple, lors du codage de réponses ouvertes concernant des caractéristiques économiques, il se peut que l'on s'écarte des procédures prescrites dans les manuels de codage et que des codes inexacts soient ainsi affectés aux différentes activités professionnelles.

8.4.6. Erreurs d'estimation

58. Les erreurs d'estimation sont dues principalement au fait qu'il n'a pas été appliqué de formule correcte pour calculer les pondérations de l'enquête. Les erreurs peuvent également être imputables à un calcul erroné des pondérations alors même que la formule correcte a été utilisée. C'est ainsi qu'il se produit des erreurs dans l'estimation de la variance d'échantillonnage (erreur d'échantillonnage) lorsque l'estimateur de la variance utilisée ne correspond pas à la conception effective de l'échantillon, ce qui entraîne des erreurs dans les intervalles de confiance liées aux estimations ponctuelles. Lorsque tel est le cas, des distorsions apparaissent dans les résultats.

8.5. Évaluation des erreurs autres que d'échantillonnage

59. Les sources d'erreurs autres que d'échantillonnage sont nombreuses et variées, comme on l'a maintes fois répété dans le présent chapitre. De ce fait, il est presque impossible d'évaluer la totalité des erreurs autres que d'échantillonnage qui surgissent lors d'une enquête. L'on peut cependant étudier et évaluer certaines des composantes d'erreurs autres que d'échantillonnage, comme on le verra ci-dessous.

8.5.1. Vérifications de la cohérence

60. Il faut, lors de l'élaboration des instruments à utiliser pour l'enquête, c'est-à-dire des questionnaires, veiller particulièrement à inclure certaines informations accessoires qui permettront de vérifier

la qualité des données rassemblées. Si ces informations supplémentaires sont faciles à obtenir, elles peuvent être rassemblées pour toutes les unités visées par l'enquête; si tel n'est pas le cas, l'on pourra se borner à les obtenir pour un sous-groupe d'unités d'échantillonnage seulement.

61. Par exemple, dans le cas d'une énumération post-recensement fondée sur la méthode *de jure,* il peut être utile de rassembler des informations sur une base *de facto* aussi pour pouvoir ainsi calculer le nombre de personnes temporairement présentes et le nombre de personnes temporairement absentes. En comparant ces deux chiffres, l'on pourra avoir une idée de la qualité des données. De même, inclure des questions débouchant sur certains ratios relativement stables, comme les ratios de masculinité, peut être utile pour évaluer la qualité des données d'enquête.

62. La cohérence devra être vérifiée aussi au stade du traitement des données. Des vérifications croisées peuvent être faites pour veiller, par exemple, à ce que les personnes codées comme étant chefs de ménage aient au moins l'âge spécifié, ou que les femmes ayant accouché n'aient pas, par exemple, moins de 13 ans.

8.5.2. Contrôle/vérification de l'échantillon

63. Pour évaluer et maîtriser certains types d'erreurs autres que d'échantillonnage, l'on peut notamment faire le travail deux fois à différentes étapes pour pouvoir détecter et rectifier plus facilement les erreurs. Pour des raisons pratiques, cette double vérification peut n'être réalisée que pour une partie du travail en ayant recours à un groupe restreint d'agents dûment formés et expérimentés. Si l'échantillon est conçu comme il convient et si l'opération de contrôle est réalisée efficacement, l'on peut non seulement détecter la présence d'erreurs mais aussi avoir une idée de leur ampleur. S'il était possible de vérifier l'intégralité du travail d'enquête, la qualité du résultat final pourrait se trouver considérablement améliorée.

64. S'agissant de la vérification de l'échantillon, les seules erreurs pouvant être vérifiées sont celles qui concernent l'échantillon vérifié. L'impact de cette contrainte peut être quelque peu atténué en divisant les produits des différentes phases de l'enquête — c'est-à-dire les listes remplies, les listes codées, les feuilles de calcul, etc. — en plusieurs lots et en procédant dans chaque lot à des vérifications par sondage. Lorsque le taux d'erreur d'un lot déterminé est supérieur au niveau spécifié, il pourra être nécessaire de vérifier l'ensemble du lot et de corriger les erreurs qui s'y trouvent afin d'améliorer ainsi la qualité du résultat final.

8.5.3. Vérifications postérieures à l'enquête ou nouvelles entrevues

65. Une vérification importante de l'échantillon, qui peut être utilisée pour évaluer les erreurs de réponse, consiste à sélectionner un sous-échantillon, ou un échantillon dans le cas d'un recensement, et à procéder à une nouvelle énumération en utilisant du personnel mieux formé et plus expérimenté que les enquêteurs initialement recrutés. Pour que cette approche soit efficace, il faut veiller à ce que :

- Une nouvelle énumération soit réalisée immédiatement après l'enquête principale pour minimiser les erreurs de mémoire;
- Des mesures soient adoptées pour minimiser *l'effet de conditionnement* que l'enquête principale peut avoir sur le travail postérieur de vérification.

66. Habituellement, l'enquête de vérification est conçue de manière à faciliter l'évaluation des *erreurs* aussi bien de *couverture* que de *contenu*. Pour cela, il est bon de commencer par recenser à nouveau toutes les unités de l'échantillon au niveau le plus général, c'est-à-dire au niveau des *zones d'énumération* et des villages, en vue de détecter les erreurs de couverture et de répéter ensuite l'enquête seulement pour un échantillon des unités finalement sélectionnées pour garantir ainsi une représentation appropriée des différents secteurs de la population qui sont particulièrement importants du point de vue des erreurs autres que d'échantillonnage.

67. L'enquête de vérification a un avantage particulier en ce sens qu'elle facilite une nouvelle vérification unitaire, qui consiste tout d'abord à comparer les données obtenues lors des deux énumérations pour les unités faisant partie de l'échantillon vérifié et à analyser ensuite les différences constatées. Lorsque des écarts apparaissent, il faut s'efforcer d'en identifier la cause et la nature et le type d'erreurs autres que d'échantillonnage.

68. S'il ne peut pas être organisé de vérification unitaire, faute de temps ou de ressources financières, une autre procédure, moins efficace cependant, appelée vérification globale, peut être utilisée. Cette méthode consiste à comparer les estimations des paramètres données par l'enquête de vérification à celles provenant de l'enquête principale. La vérification globale ne donne qu'une idée de l'erreur nette, qui est la résultante des erreurs positives et négatives. La vérification unitaire, en revanche, fournit des informations concernant les erreurs aussi bien nettes que brutes.

69. Il faut, lors de l'enquête de vérification, utiliser les mêmes concepts et les mêmes définitions que pour l'enquête initiale.

8.5.4. Méthodes de contrôle de la qualité

70. Il est éminemment possible d'appliquer des méthodes statistiques de contrôle de la qualité aux enquêtes en raison de l'ampleur et du caractère répétitif des opérations que suppose ce travail. Des listes de contrôle et des méthodes de sondage de l'acceptation peuvent être utilisées pour évaluer la qualité des données et améliorer l'exactitude du résultat final d'enquêtes de grande envergure. À titre d'exemple, 100 % du travail de chaque agent chargé de l'entrée des données peut être vérifié pendant une période initiale mais, si le taux d'erreur est inférieur à un niveau spécifié, l'on peut ensuite se borner à vérifier l'exactitude d'une partie seulement du travail accompli.

8.5.5. Étude des erreurs de mémoire

71. Les erreurs de réponse, comme indiqué ci-dessus, sont dues à plusieurs éléments, comme :

- L'attitude du déclarant à l'égard de l'enquête;
- La méthode d'entrevue;
- L'habilité de l'enquêteur;
- L'erreur de mémoire.

72. L'*erreur de mémoire* doit retenir particulièrement l'attention étant donné qu'elle présente des problèmes particuliers qui échappent fréquemment à la volonté du déclarant et qui sont liés à la longueur de la période sur laquelle porte la question et à l'intervalle entre cette période et la date de l'enquête. Ce dernier problème peut être résolu en sélectionnant une période reflétant un intervalle approprié avant la date de l'enquête ou une période aussi proche de cet intervalle que possible.

73. Pour étudier l'erreur de mémoire, l'on peut notamment rassembler ou analyser des données portant sur plusieurs périodes parmi un échantillon ou un sous-échantillon des unités sélectionnées. La principale difficulté que soulève cette approche est l'apparition d'un certain *effet de conditionnement*, qui peut être dû au fait que les données déclarées pour une période déterminée peuvent influer sur celles qui sont déclarées pour une autre période. Pour éliminer cet effet de conditionnement, l'on peut rassembler des données concernant les différentes périodes considérées parmi des unités d'échantillonnage différentes. Il y a lieu de noter à ce propos que des échantillons de grande taille sont nécessaires pour pouvoir procéder à cette comparaison.

74. Une autre méthode consiste à rassembler des informations supplémentaires qui permettent d'établir des estimations pour les différentes périodes considérées. Par exemple, dans le cas d'une enquête démographique, l'on peut rassembler des données concernant non seulement l'âge du déclarant, mais aussi le jour, le mois et l'année de la naissance. Tout écart mettra en lumière, le cas échéant, les erreurs de mémoire pouvant affecter l'âge déclaré.

8.5.6. Interpénétration des sous-échantillons

75. La méthode d'interpénétration des sous-échantillons consiste à tirer de l'ensemble de l'échantillon deux ou plusieurs sous-échantillons sélectionnés selon des procédures identiques et pouvant produire chacun une estimation valable du paramètre visé. Cette méthode aide à donner une idée de la qualité de l'information étant donné que les sous-échantillons à interpénétration peuvent être utilisés pour rassembler des données concernant les erreurs autres que d'échantillonnage comme les différences découlant du biais de l'enquêteur, des diverses méthodes utilisées pour obtenir des renseignements, etc.

76. Une fois que les sous-échantillons ont été analysés par différents groupes d'enquêteurs et traités par différentes équipes lors de la mise en tableau, une comparaison des estimations tirées des sous-échantillons permettra d'obtenir les indications générales de la qualité des résultats de l'enquête. Par exemple, si une comparaison des estimations tirée de quatre sous-échantillons traités par des équipes différentes fait apparaître trois estimations proches et une quatrième qui s'en écarte beaucoup et si cette différence est plus marquée que celle qui peut raisonnablement être imputée à une erreur d'échantillonnage, l'on peut douter de la qualité du travail effectué sur le sous-échantillon excentrique.

8.6. Conclusions

77. Il faut tenir dûment compte des erreurs autres que d'échantillonnage dans le contexte de l'enquête par sondage sur les ménages car, si elles ne sont pas réduites, elles peuvent entraîner une très forte distorsion des résultats. La plupart des enquêtes n'accordent que très peu d'attention à ces dernières, au risque de produire des résultats qui peuvent être peu fiables. Pour réduire les erreurs autres que d'échantillonnage, le meilleur moyen consiste à suivre les procédures appropriées à toutes les étapes du travail d'enquête, de la planification et de la sélection de l'échantillon jusqu'à l'analyse des résultats. Il faut en particulier veiller à ce que le personnel de terrain soit dûment formé et à mettre à l'essai les questions à poser, surtout celles qui n'ont pas été validées lors d'enquêtes précédentes.

Références et autres lectures

Biemer, P. et L. Lyberg (2003). *Introduction to Survey Quality.* Wiley Series in Survey Methodology. Hoboken, New Jersey, Wiley.

Biemer, P. *et al.,* eds. (1991). *Measurement Errors in Surveys.* Wiley Series in Probability and Mathematical Statistics. New York, Wiley.

Cochran, W. (1963). *Sampling Techniques*. New York, Wiley.

Groves, R. et M. Couper (1995). Theoretical motivation for post-survey non-response adjustment in household surveys, *Journal of Official Statistics,* vol. 11, n° 1, p. 93-106.

Groves, R. *et al.,* eds. (2000). *Survey Non-response*, Wiley-Interscience Publication. New York, John Wiley & Sons, Inc.

Hansen M., W. Hurwitz et M. Bershad (2003). Measurement Errors in Censuses and Surveys. *Landmark Papers in Survey Statistics*, IASS Jubilee Commemorative Volume.

Kalton, G. et S. Heeringa (2003). *Leslie Kish: Selected Papers.* Wiley Series in Survey Methodology. Hoboken, New Jersey, Wiley.

Kish, L. (1965), *Survey Sampling.* New York, Wiley.

Murthy, M. (1967). *Sampling Theory and Methods.* Calcutta, Inde, Statistical Publishing Society.

Onsembe, Jason (2003). Improving data quality in the 2000 round of population and housing censuses. Addis-Abeba (Éthiopie), FNUAP, Country Technical Services Team.

Organisation des Nations Unies (1982). Programme de mise en place de dispositifs nationaux d'enquête sur les ménages : *Nonsampling errors in household surveys: Sources, assessment and control.* DP/UN/INT-81-041/2. New York, Organisation des Nations Unies, Division de statistique.

Raj, D. (1972). *The Design of Sample Surveys*. New York, McGraw-Hill Book Company.

P. Chandhok (1998). *Sample Survey Theory.* Londres, Narosa Publishing House.

Shyam, U. (2004). Enquête de 2003 sur les conditions de vie au Turkménistan, Rapport technique, Institut national de la statistique et de l'information. Ashgabad (Turkménistan).

Som, R. (1996). *Practical Sampling Techniques.* New York, Marcel Dekker Inc.

Sukhatme, P. *et al.* (1984). *Sampling Theory of Surveys with Applications.* Ames, Iowa, et New Delhi, Iowa State University Press et Indian Society of Agricultural Statistics.

Verma, V. (1991). *Sampling Methods: Training Handbook.* Tokyo, Institut de statistique pour l'Asie et le Pacifique.

Whitfold, D. et J. Banda (2001). Post Enumeration Surveys: Are they Worth it? Organisation des Nations Unies, Colloque sur l'examen mondial de la série de recensements de la population et du logement de 2000 : Évaluation au milieu de la décennie et perspectives futures. New York, 7-10 août.

Chapitre 9

Le traitement des données dans les enquêtes sur les ménages

9.1. Introduction

1. Le présent chapitre, consacré au traitement des données dans les enquêtes nationales sur les ménages, décrit tout d'abord le cycle habituel d'enquête et analyse ensuite les préparatifs du traitement des données, élément qui doit faire partie intégrante du processus de planification de l'enquête.

2. L'informatique a progressé rapidement au cours des dernières années, ce qui a eu un impact profond sur les méthodes de conception et d'application des systèmes de traitement des donnés statistiques.

3. En ce qui concerne le matériel, le principal élément nouveau a été le passage de systèmes d'ordinateurs centraux à des ordinateurs personnels de plus en plus puissants en termes aussi bien de rapidité que de mémoire. Les ordinateurs personnels peuvent aujourd'hui accomplir des opérations statistiques allant d'enquêtes d'envergure modeste à de vastes opérations comme des recensements de la population et des enquêtes portant sur de très nombreux échantillons de ménage.

4. Parallèlement à l'évolution du matériel, la qualité des logiciels utilisés pour le traitement, l'analyse et la diffusion des données statistiques, s'est beaucoup améliorée aussi, de sorte que des démographes peuvent aujourd'hui accomplir des tâches qui étaient jadis réservées à des informaticiens.

5. Il a été lancé ces dernières années plusieurs logiciels de traitement des données statistiques d'enquête. Les avantages de chacun de ces logiciels dépendent du travail à accomplir. L'appendice au présent chapitre pourra être utile pour le choix des logiciels les mieux adaptés aux différentes étapes du traitement des données; il contient également une description de chacun des logiciels cités dans le chapitre.

9.2. Le cycle de l'enquête sur les ménages

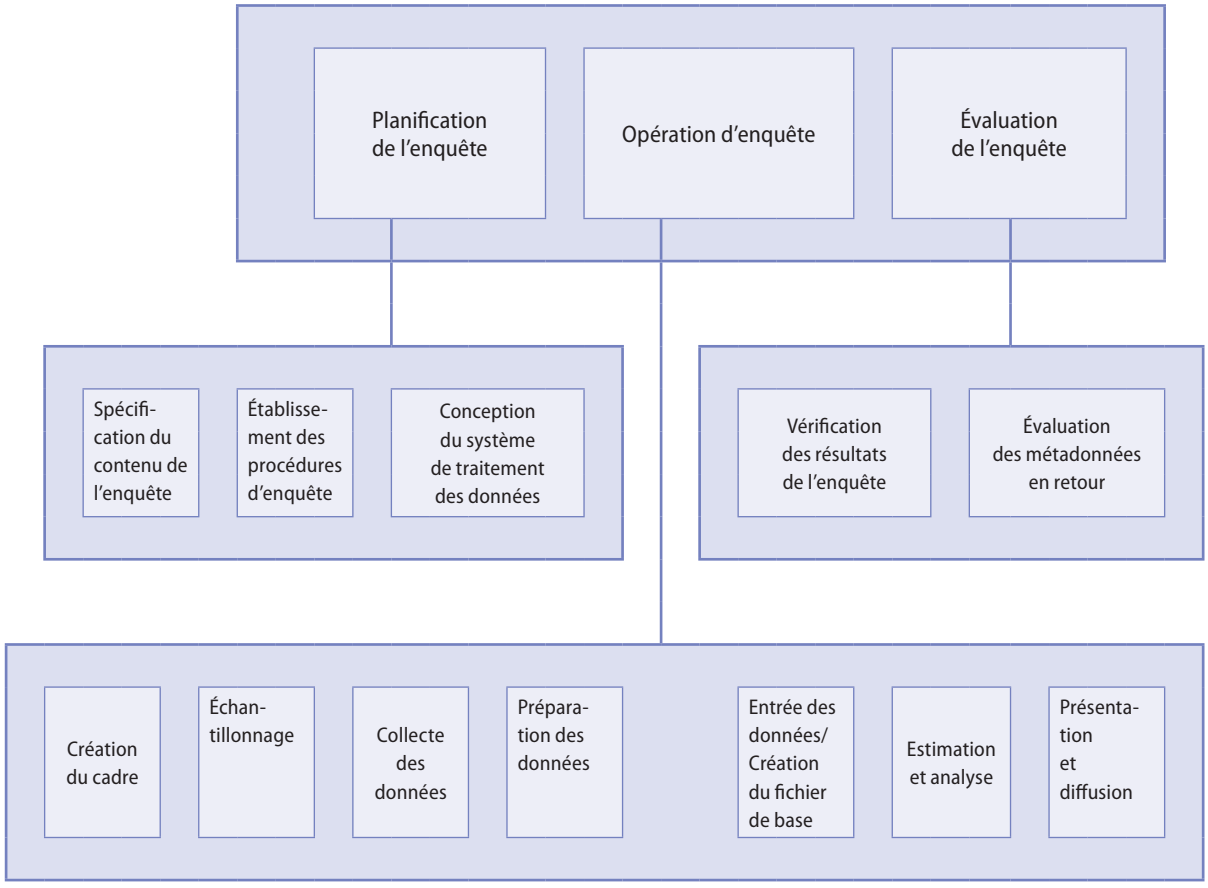
6. La figure 9.1 décrit le cycle d'enquête type. En principe, toutes les enquêtes suivent le même cycle, dont les phases sont habituellement les suivantes :

- *Planification de l'enquête :* Les concepteurs de l'enquête doivent prendre les décisions se rapportant à ses principaux objectifs et aux utilisations qui seront faites de ses résultats, aux produits attendus et aux informations à rassembler, aux procédures à suivre pour recueillir les données (conception et préparation du questionnaire et des instruments d'en-

quête connexes) et à leur transformation en produits ainsi qu'à la conception du système de traitement des données et de documentation.

- *Opérations d'enquête :* Il s'agit de créer le cadre d'échantillonnage, de concevoir et de sélectionner l'échantillon, de rassembler les données (mesure), de préparer les données (entrée, codage, édition et imputation des données), de créer le fichier d'observation (données brutes de base), de procéder aux estimations et notamment de calculer les pondérations, de créer les variables dérivées, d'analyser les données, de présenter et de diffuser les résultats.
- *Évaluation de l'enquête :* À ce stade, il s'agit de vérifier les données et de déterminer si les produits attendus ont été obtenus, si les produits ont été dûment publiés et diffusés, si les métadonnées ont été documentées et stockées, etc.

Figure 9.1.
Le cycle de l'enquête sur les ménages



Source : Sundgren (1991).

7. Avant d'entreprendre de concevoir et de mettre en place le système de traitement des données provenant d'une enquête, il importe de se représenter ce que sera l'ensemble du système. Il importe également d'échelonner comme il convient les opérations et processus. Ainsi, les objectifs de l'enquête devront déterminer la conception du produit (par exemple le plan de tabulation et les bases de données), lequel, à son tour, déterminera les activités à réaliser par la suite: conception de l'enquête, collecte, préparation et traitement des données et, finalement, analyse et diffusion des résultats.

8. Le traitement des données peut être considéré comme le processus consistant à transformer les réponses obtenues lors de la collecte des données en informations qui se prêtent à une tabulation et à une analyse. Le traitement des données englobe des activités automatisées mais aussi manuelles. Le processus peut être long et exiger beaucoup de ressources, et le traitement des données influe directement sur la qualité et le coût du produit final.

9.3. Planification de l'enquête et système de traitement des données

9.3.1. Objectifs et contenu de l'enquête

9. Comme on l'a vu au chapitre 2, il faut, lors de la conception d'une enquête, commencer par en articuler et en documenter les principaux objectifs. Les enquêtes sur les ménages ont pour but de rassembler des informations au sujet des ménages qui font partie de la population afin de répondre aux questions que les parties prenantes peuvent avoir à poser au sujet de la population cible. Comme les objectifs d'une enquête sont reflétés dans les efforts qui sont faits pour obtenir des réponses à de telles questions, le questionnaire doit être élaboré de manière à pouvoir obtenir les données pertinentes.

10. Généralement, les questions sur lesquelles les parties prenantes souhaitent travailler en se fondant sur les données recueillies au moyen d'une enquête sur les ménages peuvent être classées en plusieurs catégories (voir Glewwe, 2003).

11. Une série de questions tend à déterminer quelles sont les caractéristiques fondamentales de la population à l'étude (*proportion de la population qui vit dans la pauvreté, taux de chômage, etc.*).

12. Une autre série de questions a pour but d'évaluer l'impact des interventions des pouvoirs publics ou du progrès en général sur les caractéristiques des ménages (par exemple *la proportion des ménages qui participent à un programme déterminé, leurs caractéristiques en comparaison de celles des ménages qui ne participent pas au programme, l'amélioration ou la dégradation progressive des conditions de vie des ménages, etc.*).

13. Enfin, il y une catégorie de questions qui se rapportent aux éléments déterminants ou aux corrélations entre les circonstances et les caractéristiques des ménages (c'est-à-dire les *questions de savoir ce qui se passe et pourquoi les choses se passent ainsi*).

9.3.2. Procédures et instruments d'enquête

9.3.2.1. Plans de tabulation et produits attendus

14. Une technique utile qui aide le concepteur à établir des informations de la précision requise par l'utilisateur consiste à établir des plans de tabulation et des tableaux fictifs, qui sont des projets de tableaux qui comportent tous les éléments sauf les données proprement dites. Au minimum, le plan de tabulation doit indiquer les titres des tableaux et des colonnes et identifier les variables à tabuler, les variables de référence devant être utilisées à des fins de classification et les groupes de population

(objets de l'enquête, éléments ou unités) auxquels s'appliquent les divers tableaux (voir chapitre 2). Il est bon aussi d'indiquer les catégories de classification d'une manière aussi détaillée que possible, bien qu'elles puissent être ajustées ultérieurement lorsque la répartition de l'échantillon entre diverses catégories de réponses sera mieux connue.

15. Un plan de tabulation est important à divers points de vue. La préparation de tableaux fictifs permettra de déterminer si les données à rassembler donneront des tabulations utilisables. Ces tableaux feront apparaître non seulement ce qui manque, mais aussi ce qui est superflu. En outre, le temps supplémentaire investi dans l'élaboration d'un tableau fictif est habituellement plus que compensé, au stade de la tabulation des données, par le temps gagné sur la conception et la production des tableaux proprement dits.

16. Il y a également une étroite corrélation entre le plan de tabulation et la conception d'échantillonnage utilisée. Par exemple, une décomposition géographique des tableaux n'est possible que si l'échantillon est conçu de manière à permettre une telle ventilation.

17. L'Organisation des Nations Unies (1982) donne une description plus exhaustive de ce que représente un plan de tabulation et de ses divers avantages.

18. L'on a cité ci-dessus le rôle important que peut jouer un plan de tabulation en contribuant à une bonne planification de l'enquête et du système de traitement des données qui seront recueillies. Il importe d'insister cependant sur le fait qu'un plan de tabulation ne représente que le squelette de certains des produits que l'on peut attendre de l'enquête. Les enquêtes sur les ménages peuvent permettre de rassembler une masse d'informations. La série de microdonnées nettoyées peut être considérée comme le principal produit, et le plus fondamental. Cette série de microdonnées doit fréquemment être présentée aux parties prenantes sous une forme acceptable par les circuits de distribution appropriés.

9.3.2.2. Conception et impression des formulaires

19. Une fois que les objectifs de l'enquête et le plan de tabulation ont été déterminés, les questionnaires pertinents peuvent être établis. Le questionnaire joue un rôle central dans le processus d'enquête en permettant de transférer l'information de ceux qui la possèdent (les déclarants) à ceux qui en ont besoin (les usagers). C'est l'instrument qui reflète en termes opérationnels les besoins d'information des usagers et qui constitue le principal fondement des éléments entrés dans le système de traitement des données.

20. Selon Lundell (2003), la présentation matérielle du questionnaire à utiliser pendant une énumération a des incidences sur la capture des données et inversement. Si des techniques de numérisation ont été jugées appropriées pour la capture des données, des formulaires devront être spécialement conçus et ils varieront selon qu'il a été décidé d'utiliser des techniques de lecture optique ou d'entrée manuelle des données.

21. Quelle que soit la méthode utilisée pour l'entrée des données, chaque questionnaire doit pouvoir être identifié de façon distinctive. Comme une identification erronée du formulaire peut entraîner des doubles emplois et d'autres problèmes, il faut s'attacher à réduire ce risque au minimum. Des codes à barres seront manifestement la formule idéale lorsque sont utilisées des techniques de lecture optique. S'il est décidé d'entrer manuellement les données, l'identification du formulaire devra néanmoins contenir des informations comme un numéro de contrôle pour éviter des entrées incorrectes. Le code devra permettre d'identifier chaque questionnaire séparément et devra toujours

être numérique. Habituellement, les informations nécessaires pour l'affectation de pondérations ou de facteurs d'expansion (strates, unitaires primaires d'échantillonnage, segments de zone, distinctions entre circonscriptions administratives nécessaires aux fins de la tabulation, etc.) sont également jointes au formulaire.

22. Les formulaires sont habituellement reliés sous forme de cahiers (par exemple un cahier pour chaque zone d'énumération, etc.). Chaque cahier, comme les formulaires, doit porter un code d'identification propre, et il doit y avoir une corrélation clairement spécifiée entre chaque cahier et formulaire qu'il contient, pour garantir que le formulaire X appartient toujours au cahier Y et seulement au cahier Y. Les codes d'identification des cahiers seront utilisés pendant toute la phase de traitement des données, depuis l'indication de l'arrivée d'un cahier jusqu'à la recherche d'un formulaire lorsque cela est nécessaire, par exemple pour vérifier tel ou tel élément pendant la tabulation ou l'analyse des données. Il faut par conséquent réduire au minimum les risques d'identification non seulement de chaque formulaire mais aussi de chaque cahier.

23. Chaque champ doit être conçu de manière à pouvoir entrer le maximum de caractères possibles; par exemple, il faut indiquer très clairement le nombre maximum possible de membres du ménage pour pouvoir prévoir un champ de taille correcte.

24. Il importe de veiller à éliminer tout risque d'erreur en ce qui concerne la définition des unités d'observation, les questions pouvant être sautées et les autres aspects du questionnaire. Toutes les enquêtes sur les ménages ont pour but de rassembler des informations au sujet d'une unité statistique majeure (l'objet principal), à savoir le ménage, ainsi que de différents groupes subordonnés (objets associés) du ménage: personnes, rubriques budgétaires, terres de culture, récoltes, etc. Le questionnaire doit indiquer de façon claire et explicite ce que sont ces unités et ainsi faire en sorte que chaque unité observée se voie affecter une étiquette appropriée permettant de l'identifier en la distinguant de toutes les autres. Une des méthodes fréquemment utilisées pour identifier des ménages, qui constitue également un élément important lorsque les données sont entrées manuellement, consiste à utiliser un simple numéro de série écrit, tamponné ou pré-imprimé sur la page de couverture du questionnaire. Habituellement, le numéro de série représente également le code d'identification du formulaire.

25. Comme, de plus en plus, l'on utilise des techniques de lecture optique pour capturer rapidement les données, il importe aussi de tenir compte des caractéristiques spéciales que doit présenter le questionnaire pour pouvoir être numérisé. Certains des aspects à prendre en considération lorsque le questionnaire doit être traité par lecture optique — par exemple au moyen d'un logiciel de reconnaissance optique de caractères (OCR), de reconnaissance intelligente de caractères (ICR) ou de lecture optique (OMR) — sont évoqués ci-dessous.

26. Deux codes à barres sont portés sur le questionnaire. Le premier est le code qui identifie chaque page du questionnaire, ce qui est important, surtout lorsque les pages se présentent d'une façon très semblable. C'est le principal moyen qui permet au logiciel de lecture de distinguer les différentes pages du questionnaire. Un deuxième code à barres, habituellement avec l'interprétation connexe, est placé sur chaque page du questionnaire et est exactement identique sur chaque page, tout en reflétant un numéro d'ordre différent pour chaque questionnaire. Ce code à barres relie ensemble les pages d'un même questionnaire, ce qui est très important étant donné que les pages doivent habituellement être séparées pour pouvoir être numérisées.

27. L'agencement exact des champs sur le formulaire imprimé représente le dictionnaire des données devant être rassemblées. Si l'intention est de capturer un code de district d'énumération à cinq

chiffres, par exemple, il est imprimé sur le questionnaire un champ comportant cinq cases pour identifier le code du district. Si le champ de capture est conçu en vue d'une entrée manuelle des données, cette précision ne sera pas nécessaire sur le questionnaire attribué et il suffira de prévoir un champ ouvert où puissent être portés cinq chiffres. Il importe néanmoins d'affiner la conception du système d'entrée des données de sorte que le champ soit identifié par un numéro et soit automatiquement rempli pour éviter un manque d'alignement des chiffres et une interprétation erronée des données.

28. La conception du formulaire variera beaucoup aussi selon que les informations seront numérisées ou entrées manuellement étant donné que le dispositif de lecture s'en remet entièrement à la position des champs de données pour identifier les informations dont il s'agit. À la différence de l'entrée manuelle, il n'est pas nécessaire, pour une lecture optique, d'identifier les champs imprimés sur le questionnaire, indépendamment de quelques champs d'ajustement par page. Pour que le système de lecture puisse interpréter correctement l'image, les champs ne doivent être ni trop petits, ni trop grands. Les enquêteurs devront être encouragés, pendant la formation ainsi que pendant le travail de terrain, à inscrire des caractères clairs et distincts centrés dans le champ de données.

29. Il existe également une différence fondamentale en ce qui concerne la façon dont le questionnaire est censé capturer, par exemple les codes d'activités professionnelles, entre les processus manuels de codage appropriées pour une entrée manuelle des données et un codage sur écran à l'aide d'une recherche des codes assistée par ordinateur. Lorsque le questionnaire est conçu en vue d'une entrée manuelle des données, l'indication du code est imprimée sur le questionnaire à l'intention de la personne qui sera chargée du codage après avoir examiné la réponse ouverte à la question: « *Quelle est votre profession?* » Si le questionnaire doit être numérisé, il n'est pas absolument nécessaire de réserver de l'espace à cet effet sur le questionnaire imprimé car la liste de codes est intégrée à la sortie d'imprimante, de sorte que les personnes responsables du codage et de l'entrée des données se borneront à effectuer un travail de vérification. Lors de l'entrée du code désignant l'activité professionnelle, le vérificateur se trouve en présence d'une réponse ouverte sur l'image numérisée et d'un menu de codes indexés parmi lesquels le code approprié puisse être sélectionné rapidement.

30. Lorsque l'on envisage d'utiliser des techniques de lecture optique, la qualité de l'impression devient également une question importante. Les dispositifs de lecture optique sont plus sensibles que l'œil humain à des imperfections de l'impression,. Il peut surgir des problèmes, par exemple, lorsque sont utilisées certaines couleurs ou combinaisons de couleurs, lorsque la tonalité et l'acuité varient, lorsque des champs sont imprimés de biais ou mal placés, et lorsqu'il y a des erreurs dans la numérotation automatique des pages ou des erreurs de collation.

31. St. Catherine (2003) et Lundell (2003) présentent certaines des questions à prendre en considération lorsqu'il est envisagé d'utiliser des méthodes de lecture optique pour le traitement des données provenant d'enquêtes statistiques et de recensement.

9.3.3. Conception des systèmes de traitement des données dans le cadre des enquêtes sur les ménages

9.3.3.1. Approche générale du traitement des données dans le cadre des enquêtes sur les ménages

32. La conception des systèmes constitue l'un des principaux aspects de la planification d'une enquête sur les ménages. Essentiellement, à ce stade, il faut spécifier de façon formelle les données à rassembler et l'ensemble du système de traitement des données.

33. Jambwa, Parirenyatwa et Rosen (1989) présentent comme suit les avantages qu'offrent l'adoption et l'utilisation d'un système formel de conception, de mise au point et de documentation de tous les systèmes d'un office de la statistique, particulièrement dans le contexte des enquêtes sur les ménages :

a) Le système peut être le cadre de coopération requis entre les statisticiens, les spécialistes de la question traitée et les analystes de systèmes/programmeurs;

b) Toutes les opérations devant être réalisées lors de l'enquête seront explicitement décrites et documentées et l'on pourra s'y référer ultérieurement. La documentation correspondante (c'est-à-dire la série de métadonnées) sera importante tant pour la mise au point que pour la maintenance des systèmes de production de statistiques;

c) Le coût de la mise au point et de la maintenance des systèmes tend à être assez élevé étant donné les nombreux systèmes différents dont a habituellement besoin un office de statistique. La structure des enquêtes sur les ménages tend à suivre le même schéma et les mêmes principes. Par exemple, ces enquêtes partagent généralement les mêmes types de fichiers et de données, de systèmes de codage, etc. Les enquêtes suivantes peuvent par conséquent se trouver facilitées lorsque des systèmes de traitement des données ont déjà été mis au point, ce qui devrait pouvoir réduire les coûts de mise au point et de maintenance;

d) L'adoption d'une approche formelle est importante aussi pour l'intégration de l'enquête si, par exemple, l'on souhaite mener une analyse combinée de données provenant de différentes enquêtes ou de différentes séries de la même enquête.

9.3.3.2. Caractéristiques générales d'un système informel de conception des systèmes

34. L'on trouvera dans la présente section une indication de certains des aspects généraux et fondamentaux du système formalisé décrit ci-dessus.

Structure des données

35. Les décisions concernant la question sociale à analyser, les données à utiliser et la technique statistique à appliquer revêtent une importance fondamentale pour la qualité de l'analyse. Toutefois, une question encore plus fondamentale est celle de l'identification et de la définition des objets de l'enquête ou des unités d'analyse. Nous avons déjà insisté sur ce point dans la section concernant la conception et l'impression des formulaires. Lors de la mise au point du système de traitement des données qui sera utilisé, il y aura lieu de donner une description plus formelle et plus détaillée de l'unité d'analyse (objet) et des variables.

36. Selon Sundgren (1986), l'objet ou l'unité d'analyse doit être défini comme étant toute entité concrète ou abstraite (objet matériel, créature vivante, organisation, événement, etc.) à propos duquel les usagers souhaitent avoir des informations. La définition d'un objet est indissociablement liée à la question sociale (objectifs de l'enquête) à propos de laquelle des données sont rassemblées et analysées. De même, dans le cas des enquêtes sur les ménages, les objets, éléments ou unités à propos desquels les parties prenantes souhaitent avoir une information sont, par exemple, le ménage, la personne, le terrain, etc. Le plus souvent, l'objet principal est le ménage et l'objet principal est habituellement accompagné de plusieurs objets connexes, qui dépendront selon l'enquête.

37. Le tableau 9.1 présente les objets/unités définis en vue de l'Enquête démographique intercensitaire réalisée au Zimbabwe en 1987. L'objet principal est le ménage et les objets associés étaient « personne », « femme de 12 ans et plus » et « défunt » (Lagerlof, 1988).

Tableau 9.1
Exemple d'objets/unités d'analyse tiré de l'Enquête démographique intercensitaire réalisée au Zimbabwe en 1987

Objet/unité	Variables distinctives	Définition de l'objet/unité	Variables importantes	Objets connexes	
				Objet	Clé étrangère
MÉNAGE	HID = identification du ménage (zone, division, subdivision, EANR = numéro de la *zone d'énumération*, HHNR = numéro du chef de ménage)	Un foyer est un groupe de personnes qui, normalement, vivent et prennent leurs repas ensemble, et exclut les visiteurs	SOH = taille des ménages STRATE ZONE	PERSONNE DÉFUNT	HID HID
PERSONNE	HID, PID PID = identification de la personne	La personne est un membre usuel du ménage ou un visiteur qui y a passé la nuit dernière	SEXE, ÂGE, MARSTAT = situation conjugale, ETHNIQUE = groupe ethnique, USMEM = membre usuel des ménages, RELTH = relations avec le chef de ménage	MÉNAGE FEMME ≥12 ANS	HID HID, PID
DÉFUNT	HID, DID DID = identification du défunt	Le défunt était habituellement membre du ménage au cours des 12 mois écoulés.	SEXD = sexe du défunt AGED = âge du défunt	MÉNAGE	HID
Femme ≥ 12 ans	HID, DID	Toute femme qui a 12 ans ou plus ou qui est habituellement membre du ménage ou qui y a passé la nuit dernière.	Nombre d'enfants nés	PERSONNE	HID, DID

38. Pour chaque objet, il y aura plusieurs variables à étudier. Les variables sont les propriétés (attributs ou caractéristiques) des objets. Par exemple, l'objet « personne » peut avoir comme variables l'âge, le revenu, l'activité professionnelle, la situation conjugale, etc. Les variables peuvent être qualitatives ou quantitatives.

39. Chaque objet doit également avoir une identification distinctive. L'identification d'un objet associé indique l'objet fondamental auquel il se rattache. Par exemple, la « personne » serait liée au « ménage » et serait identifiée par la combinaison de l'identification du ménage (HID) et celle de la personne (PID), c'est-à-dire le numéro de série à l'intérieur du ménage (PID).

Éléments entrés dans le système de traitement des données

40. Les éléments à entrer sont les valeurs obtenues enregistrées par les enquêteurs au moyen du questionnaire.

Produit du système de traitement des données

41. Le produit du système revêt principalement la forme de tableaux statistiques (établis conformément au plan de tabulation), de bases de données contenant des micro ou des macrodonnées, etc. Ces tableaux varieront en fonction du type d'objet, du type de variable et du type de mesure statistique. Les variables présentées dans les tableaux sont habituellement « originelles » mais peuvent également être dérivées de variables originelles.

Organisation des fichiers

42. Habituellement, il conviendra d'avoir pour les fichiers des structures différentes au stade de l'entrée de données et au stade précédant la tabulation. Par exemple, un fichier de longueur variable pourra être préférable pour l'entrée des données car les ménages varient par leur taille et par leur composition, d'où la nécessité d'avoir pendant l'entrée des données des registres de longueur variable. Cette méthode utilise l'espace efficacement mais présente des inconvénients ultérieurs au stade du traitement des données. En définitive, cependant, il y a fréquemment lieu d'analyser les données en fichiers « plats » de manière à faciliter la tabulation et à pouvoir utiliser au mieux les différents types de logiciels non spécialisés.

43. *Schéma logique du système.* Il conviendra d'élaborer un schéma logique raisonnablement détaillé décrivant l'enquête. Ce schéma est important pour de nombreuses raisons. Tout d'abord, il permet d'établir des calendriers et d'estimer les ressources nécessaires pour mener à bien le traitement des données. Habituellement, les principales activités de traitement des données, quelle que soit l'enquête, sont les suivantes :

a) Vérification, édition et codage des données;

b) Entrée, vérification et validation des données;

c) Transformation de la structure des données entrées en une structure permettant des tabulations;

d) Tabulation.

44. Le schéma logique du système devra également faire apparaître les principales opérations à effectuer sur les fichiers, comme la sélection, la projection, le tri et le rapprochement des fichiers, la dérivation de nouvelles variables, l'agrégation, la tabulation et la présentation graphique.

Système de documentation

45. Il importe de constituer une documentation claire et complète (c'est-à-dire une série de métadonnées) aussi bien pour la mise au point que pour la maintenance des systèmes de traitement des données. Il faut par conséquent documenter les fichiers et les différentes opérations réalisées de sorte que les personnes qui n'ont pas pris part à la mise en œuvre du système originel puissent également les utiliser. Il conviendra, pour veiller à ce que la documentation soit suffisante, d'utiliser et de stocker sur ordinateur un modèle normalisé, accompagné de ses données.

46. Dans toute la mesure possible, il faudra utiliser les mêmes appellations, les mêmes codes et le même format pour les variables utilisées dans le système de traitement des données des différentes enquêtes réalisées par l'office de statistique si les codes ont la même signification. Cela est particulièrement important dans le cas des variables qui sont utilisées pour identifier les objets figurant dans

le fichier, ces variables pouvant également être utilisées pour combiner des données provenant de systèmes différents.

47. Des modèles d'écrans d'entrée des données et de lecture numérique des formulaires sont intégrés au logiciel pour faciliter la documentation, laquelle ne peut être efficace que s'ils sont pleinement utilisés. C'est ainsi par exemple que les formats de dictionnaires de données Census and Survey Processing System (CSPro) ou le Système intégré de traitement sur microprocesseur (IMPS) définissent la position de chaque variable dans le fichier de données — avec les points de départ et d'aboutissement — que la variable soit un chiffre ou un caractère et qu'elle soit ou non récurrente et quel que soit le nombre de fois où elle réapparaît. Ce dictionnaire étiquette également les valeurs contenues dans la variable (St. Catherine, 2003) [voir l'appendice au présent chapitre pour de plus amples informations sur les systèmes CSPro et IMPS].

9.4. Opérations d'enquête et traitement des données

9.4.1. Création du cadre et conception de l'échantillon

48. Comme on l'a vu aux chapitres 3 et 4, les unités primaires d'échantillonnage, dans le cas de nombre d'enquêtes sur les ménages, sont les zones d'énumération définies en vue du dernier recensement national disponible. Créer un fichier contenant la liste de toutes les zones d'énumération du pays est un bon moyen, communément utilisé, d'élaborer le cadre d'échantillonnage primaire, et le mieux est d'utiliser à cette fin un programme comme Microsoft Excel, avec une ligne pour chaque *zone d'énumération* et des colonnes pour toutes les informations devant être réunies (voir l'appendice au présent chapitre pour plus amples informations sur le système Excel).

49. Le cadre doit être d'accès et d'utilisation faciles pour différentes manipulations comme tri, filtrage et production de statistiques sommaires pouvant faciliter la conception de l'échantillon est les estimations. Le système Microsoft Excel est facile à utiliser, et il comporte des fonctions de tri, de filtrage et d'agrégation qui sont indispensables pour préparer les échantillons à partir du cadre d'échantillonnage. Les tableaux peuvent également être importés dans la plupart des autres logiciels. Il est généralement plus commode de créer un tableau différent pour chaque strate d'échantillonnage.

50. Le contenu des unités du cadre d'échantillonnage devra être le suivant :

- Un code d'identification primaire, numérique, qui distingue toutes les divisions et subdivisions administratives dans lesquelles se trouve l'unité considérée. Il sera bon de numéroter les unités dans l'ordre géographique. Habituellement, les codes des *zones d'énumération* présentent déjà ces attributs.
- Un code d'identification secondaire, qui sera le nom du village (ou de la circonscription administrative considérée) où se trouve l'unité, qui servira à localiser celle-ci sur une carte et sur le terrain.
- Un certain nombre de caractéristiques de l'unité d'échantillonnage, comme la taille (population, ménages), le milieu urbain ou rural, la densité de population, etc., caractéristiques qui peuvent être utilisées pour une stratification ou pour l'affectation de probabilités de sélection, et comme variables auxiliaires pour les estimations.
- Des données opérationnelles comme des informations sur les modifications apportées aux unités et une indication de l'usage fait de l'échantillon.

51. Les procédures de sélection et les probabilités de sélection pour toutes les unités d'échantillonnage doivent être pleinement documentées pour toutes les phases. Lorsque des échantillons-maîtres sont utilisés, il devra être indiqué ceux qui auront été employés pour constituer les échantillons destinés aux différentes enquêtes. Un système standard de numéros d'identification doit être utilisé pour les unités d'échantillonnage.

52. L'échantillon-maître élaboré par la Namibie sur la base du Recensement de la population et du logement de 1991 peut servir d'exemple de ce qui a été dit ci-dessus (Office central de statistique, Namibie, 1996).

Exemple :

Pour pouvoir sélectionner un échantillon aléatoire de zones géographiques en Namibie, il a fallu créer un cadre d'échantillonnage de ces zones, constituant les unités primaires d'échantillonnage (UPE). Ces zones contenaient en moyenne une centaine de ménages, la plupart d'entre elles de 80 à 100. Elles étaient fondées sur les *zones d'énumération* utilisées lors du Recensement de la population et du logement de 1991. Les *zones d'énumération* les plus petites étaient combinées aux *zones* adjacentes pour constituer des UPE de taille suffisante, la règle générale étant que chaque UPE devrait englober au moins 80 ménages. En tout, il a été constitué environ 1 685 UPE, regroupées en strates par région ainsi que par milieu rural, par petites agglomérations et par milieu urbain.

53. La stratification a été fondée sur une classification des *zones d'énumération* réalisée lors des préparatifs du Recensement de 1991. En tout, il a été créé 32 strates à l'intérieur desquelles les UPE ont été rangées dans l'ordre géographique. En milieu urbain et dans les petites agglomérations, les UPE ont également été rangées en fonction du niveau de revenu, les zones à revenu élevé venant en premier, suivies par les zones à revenu intermédiaire ou à faible revenu.

54. L'Office central de statistique a préparé des fichiers Microsoft Excel sur la base du cadre d'échantillonnage, lequel contenait les indications suivantes :

- Région;
- Code d'identification de l'UPE;
- Niveau de revenu (en milieu urbain seulement);
- District;
- Numéro(s) des *zones d'énumération;*
- Nombre de ménages selon le Recensement de 1991;
- Nombre cumulatif de ménages par strates;
- Population, par sexe, selon le Recensement;
- Situation en regard de l'échantillon-maître (l'UPE faisant ou non partie de l'échantillon-maître);
- Numéro de l'UPE dans l'échantillon-maître (seulement pour les UPE faisant partie de l'échantillon-maître);
- Pondérations (facteurs d'augmentation ou d'inflation), seulement pour les UPE faisant partie de l'échantillon-maître.

55. Il a été constitué un fichier Microsoft Excel pour chaque région et, à l'intérieur de chaque fichier, les UPE ont été regroupées selon qu'elles appartenaient à un milieu rural, à de petites agglo-

mérations, aux tranches de revenu élevé en milieu urbain et aux tranches de revenus intermédiaires et de faibles revenus en milieu urbain et dans les petites agglomérations.

56. L'on trouvera dans Pettersson (2003) une analyse détaillée des questions liées aux échantillons-maîtres. Munoz (2003) indique en outre comment un cadre informatisé comme celui qui est décrit ci-dessus peut également être employé pour appliquer la procédure d'échantillonnage en la guidant tout au long de ses différentes phases: organisation du cadre primaire, habituellement fondé sur les résultats du dernier recensement de la population et du logement, sélection des unités primaires d'échantillonnage sur la base d'une probabilité proportionnelle à la taille (taille mesurée par le nombre de ménages, de logements ou d'habitants), actualisation du tableau sur la base de la liste de ménages sélectionnés et calcul des probabilités de sélection et des pondérations d'échantillonnage correspondantes. La section 9.4.3.5 ci-dessous, concernant les procédures d'estimation ponctuelles et le calcul des pondérations, contient une analyse plus détaillée du calcul des probabilités de sélection et des pondérations correspondantes (voir également le chapitre 5). Les données requises pour ces calculs peuvent être tirées d'un tableau, comme indiqué ci-dessus et les pondérations peuvent être calculées en utilisant le tableau, comme démontré par Munoz.

9.4.2. Collecte et gestion des données

57. Les enquêtes sur les ménages ont débouché sur un très grand nombre de questionnaires remplis. Il faut, si l'on veut éviter le chaos, bien réfléchir d'entrée et définir sans tarder les procédures qui seront suivies pour le traitement matériel et la comptabilisation de ces masses de documents. Les méthodes de traitement manuel (enregistrement et recherche) des questionnaires doivent être soigneusement planifiées et doivent être opérationnelles bien avant que les données provenant du terrain ne commencent d'arriver. Un aspect important de ce système est l'estimation du volume des données attendues, de manière à pouvoir se procurer les dossiers, cartons, etc., nécessaires et réserver de l'espace sur des étagères ou dans des placards. Le deuxième élément du système est un registre où des informations concernant les questionnaires peuvent être portées lors de leur arrivée et où l'on puisse suivre le cheminement des données tout au long du système. Ce sont là des aspects clés de la gestion des données et des préalables importants au succès de la gestion et de la mise en œuvre de la stratégie de traitement des données.

58. Il importe également, à ce stade, de prévoir comment sera garantie la sécurité matérielle des questionnaires remplis. C'est là un des domaines dans lesquels la lecture optique constitue une formule attrayante. En effet, si les questionnaires sont numérisés dès leur arrivée, le risque de perte des données figurant sur les questionnaires se trouvera réduit. La lecture optique offre un degré de sécurité supplémentaire en permettant de sauvegarder les questionnaires numérisés (Edwin, 2003). Il importe de noter toutefois que les résultats dépendront directement de la façon dont cette méthode et les processus connexes auront été organisés et gérés par l'institution responsable. La lecture optique a donné de bons résultats dans certains pays mais pas dans d'autres. Pour qu'elle soit utile, il faut, entre autres, établir dans quelle mesure l'office de statistique utilise des procédures centralisées ou décentralisées, quel est le profil des enquêteurs et quelles sont les garanties de qualité qu'offrent les instruments de collecte de données.

9.4.3. Préparation des données

59. Les données rassemblées doivent être entrées dans un fichier. La traduction des données figurant sur les questionnaires en données informatisées est ce que l'on appelle l'entrée des données. Dans ce contexte, il faut souvent ranger en différentes catégories des valeurs variables provenant de réponses ouvertes; ce processus de classement est appelé codage. En éditant les données obtenues, l'on peut identifier celles qui sont erronées et adopter les mesures appropriées pour vérifier les erreurs soupçonnées, par exemple en interrogeant à nouveau la source de l'information. Ces vérifications peuvent être suivies d'une actualisation (correction). Les différentes étapes du processus sont notamment l'entrée, le codage, l'édition, la vérification et l'actualisation/correction des données. Collectivement, ces opérations sont appelées ici préparation des données.

9.4.3.1. Stratégies de préparation des données

60. Munoz (2003) expose en détail les différents aspects et les configurations des opérations de préparation des données. La formule la plus communément utilisée dans le cas des enquêtes sur les ménages consiste à centraliser la préparation des données après leur collecte sur le terrain. Une autre formule peut consister à intégrer l'entrée des données aux opérations sur le terrain. L'innovation la plus récente est la technique d'entrevue assistée par ordinateur.

Préparation centralisée des données

61. Telle était la seule option avant l'arrivée des ordinateurs personnels et cette formule est encore aujourd'hui celle qui est la plus utilisée pour les enquêtes dans les pays en développement, avec de légères modifications rendues possibles par l'introduction des micro-ordinateurs. Selon cette approche, l'entrée des données est considérée comme un processus industriel à mener à bien après les entrevues en un endroit centralisé ou en plusieurs endroits, comme le siège ou les bureaux régionaux de l'office national de statistique.

Préparation des données sur le terrain

62. Dernièrement, l'intégration de contrôles informatisés de la qualité des opérations sur le terrain a été l'un des principaux moyens auxquels l'on a eu recours pour améliorer la qualité et l'actualité des enquêtes sur les ménages. Selon cette stratégie, l'entrée des données et les contrôles de cohérence font partie intégrante des opérations sur le terrain.

63. À cette fin, le préposé à l'entrée des données peut être appelé à travailler sur un ordinateur de bureau dans une localité fixe (par exemple le bureau régional de l'office national de statistique) et à organiser le travail sur le terrain de sorte que les autres membres de l'équipe visitent chaque lieu d'enquête (généralement une unité primaire d'échantillonnage) au moins deux fois de manière à ménager le temps nécessaire pour entrer les données et en vérifier la cohérence entre deux visites. Lors de la deuxième visite et des visites ultérieures, l'enquêteur devra poser à nouveau aux ménages intéressés les questions dans le cas desquelles il a été détecté des erreurs, des omissions ou des contradictions lors de l'entrée des données.

64. Selon une autre méthode, le préposé à l'entrée des données travaille sur un ordinateur portable et se joint aux autres membres de l'équipe lors de leurs visites sur le terrain. L'ensemble de l'équipe

reste sur place jusqu'à ce que toutes les données soient entrées et soient acceptées comme étant complètes et correctes par le programme d'entrée de données.

65. Les avantages de la méthode consistant à intégrer la collecte et la préparation des données sont notamment la possibilité d'obtenir des informations de meilleure qualité étant donné que les erreurs peuvent être corrigées tandis que les enquêteurs se trouvent encore sur le terrain, qu'il est possible de générer des bases de données et d'entreprendre la tabulation et l'analyse des données peu après l'achèvement des opérations sur le terrain et qu'il est plus facile de normaliser la collecte des données par les enquêteurs.

66. Quelle que soit celle des deux approches susmentionnées qui sera retenue, il importe au plus haut point d'assurer une alimentation constante en électricité là où sont menées les opérations. Dans les pays où le réseau électrique est peu développé, comme dans la plupart des pays en développement, surtout en milieu rural, de telles formules ne seraient tout simplement pas viables. Il convient d'ajouter que l'utilisation de matériel mobile pour la collecte et la préparation des données suscite différents problèmes d'organisation et de logistique. Pour pouvoir utiliser cette stratégie avec succès, il faut disposer d'un système de gestion efficace, garantir la sécurité du matériel et des données, disposer de systèmes adéquats de sauvegarde des données et prévoir un approvisionnement en quantités suffisantes de matériel consomptible, comme des accumulateurs et des piles de rechange.

Entrevue assistée par ordinateur

67. L'entrevue personnelle assistée par ordinateur est une forme d'entrevue personnelle selon laquelle l'enquêteur, plutôt que de remplir un questionnaire sur le papier, utilise un ordinateur portable pour entrer directement les informations rassemblées dans la base de données. Cette méthode permet de gagner du temps lors du traitement des données et évite à l'enquêteur de devoir transporter avec lui des centaines de questionnaires. Cependant, bien que cette technologie existe depuis de nombreuses années, très peu de choses ont été faites pour appliquer sérieusement cette stratégie à des enquêtes complexes dans les pays en développement. Cette méthode de collecte de données peut être onéreuse et l'enquêteur, pour l'utiliser, doit être familiarisé avec l'informatique et doit savoir dactylographier. Les entrevues assistées par ordinateur exigent également une bonne préparation, l'entrevue devant avoir un début et une fin bien définies. Toutefois, la plupart des enquêtes réalisées dans les pays en développement exigent de multiples visites de chaque ménage, des entrevues séparées avec chaque membre du ménage, etc., selon un processus qui n'est pas rigoureusement structuré mais qui, pour l'essentiel, est plutôt dirigé par l'enquêteur.

9.4.3.2. Codage et édition des données

68. La vérification, l'édition et le codage des données constituent probablement la phase la plus difficile du traitement des données. C'est lors de l'organisation de la gestion et de la préparation des données que les enquêteurs fraîchement émoulus se heurtent souvent aux plus grandes difficultés. Si possible, les processus de vérification, d'édition et de codage ont intérêt à être informatisés. Cependant, dans le cas du codage, il faudra naturellement tenir compte des cas dans lesquels des codes ne peuvent pas être affectés automatiquement, auquel cas il faudra procéder manuellement.

Codage

43. L'objectif est de préparer des données de manière qu'elles soient prêtes à être entrées ans l'ordinateur. L'opération de codage consiste essentiellement à affecter des codes numériques aux réponses enregistrées verbalement (par exemple concernant la localité géographique, la profession, la branche d'activité, etc.). Elle peut également consister à transcrire les données, auquel cas les codes numériques déjà affectés et enregistrés pendant l'entrevue sont reportés sur des feuilles de codage.

44. Il y aura lieu de préparer un manuel contenant des indications explicites à l'intention des préposés au codage. Ce manuel devra contenir une série de catégories couvrant toutes les réponses acceptables aux questions posées. Dans le cas d'une enquête sur les ménages de grande envergure, il y aura lieu de s'attacher, dans tous les cas où cela sera possible, à rédiger les questions de sorte que les réponses soient fermées et puissent être pré-codées.

Édition et vérification des données

45. La vérification et/ou l'édition des questionnaires a pour but : *a*) de veiller à la cohérence des données et à la cohérence des tableaux et entre les tableaux; et *b*) de détecter, vérifier, corriger et éliminer les valeurs périphériques, étant donné que des valeurs extrêmes contribuent beaucoup à la variabilité des estimations.

46. Le processus d'édition consiste à réviser ou corriger les mentions figurant dans les questionnaires. Il s'agit en quelque sorte d'une procédure de validation et les impossibilités de données, ou d'une procédure statistique, des vérifications étant opérées sur la base d'une analyse statistique des données. De plus en plus, le travail d'édition est fait par ordinateur, soit au stade de l'entrée des données, soit au moyen de vérifications, interactives ou non, selon que l'opérateur peut ou non corriger et éditer les erreurs détectées. Pour rectifier les erreurs plus complexes, cependant, il faudra plus de temps et des analyses plus approfondies avant de pouvoir trouver la correction appropriée et, en pareil cas, des vérifications non interactives sont préférables. Les auteurs cités par Olsson (1990) donnent des informations détaillées sur les divers aspects de la vérification et de l'édition des données d'enquête.

73. *Vérification et édition manuelles.* La vérification ou l'édition manuelle a essentielle pour but de détecter les omissions, contradictions et autres erreurs évidentes dans les questionnaires avant que les données ne commencent à être traitées. L'édition manuelle doit commencer dès que possible et aussi près que possible de la source de données, par exemple au bureau provincial, de district ou local. Idéalement, la plupart des erreurs que contiennent des données peuvent être détectées et corrigées sur le terrain avant que les formulaires soient envoyés au bureau central pour dépouillement. Ainsi, lors de la formation initiale et dans le manuel d'instruction, l'enquêteur et son superviseur sont habituellement invités à vérifier les questionnaires et à corriger les erreurs éventuelles tandis qu'ils se trouvent encore sur le terrain, avant de renvoyer les données. Il s'agit là d'une tâche importante et difficile qui, pour être menée à bien, exige un travail de terrain de qualité, une supervision et une gestion efficaces de l'enquête, etc.

74. *Édition assistée par ordinateur*. L'on peut éditer les données avec l'aide d'un ordinateur : *a*) de façon interactive au stade de l'entrée des données; *b*) en utilisant un traitement par lots après l'entrée des données; et *c*) en combinant les méthodes *a* et *b.* L'édition interactive est généralement la plus utile dans le cas d'erreurs simples (par exemple des erreurs de saisie), et elle retarderait le processus de capture des données dans le cas d'erreurs qui exigent de consulter les superviseurs. La rectifica-

tion de ces erreurs, y compris celles qui sont dues à la non-réponse, doit faire l'objet d'une opération d'édition distincte.

75 Les programmes d'édition assistée par ordinateur sont souvent conçus à partir de plates-formes comme le Système intégré de traitement sur microprocesseur (IMPS), le Système intégré pour l'analyse des enquêtes (ISSA), le Census and Survey Processing System (CSPro), Visual Basic et Microsoft Access (l'on trouvera de plus amples détails sur ces logiciels dans l'appendice au présent chapitre). Les programmes les plus simples filtrent les données, entrée par entrée, et relèvent les contradictions sur la base des règles d'édition incorporées aux programmes. Dans le cas des programmes plus poussés, les variables (par exemple les variables d'identification) que contiennent les différents fichiers peuvent être comparées, et les contradictions signalées. Ces systèmes produisent des listes d'erreurs, qui sont alors comparées normalement aux données brutes. Les erreurs sont corrigées dans un des exemplaires du fichier de données brutes.

Types de vérifications

76. Les données que contiennent les questionnaires doivent faire l'objet de divers types de vérifications, qui portent notamment sur la complétude des données, les données de référence, les questions sautées, la cohérence des informations et l'exactitude de l'entrée des données (Munoz, 2003).

77. *Vérifications de complétude*. Les vérifications de complétude permettent de s'assurer que chaque variable ne contient que des données se trouvant à l'intérieur d'un domaine limité de valeurs valides. Les variables catégoriques ne peuvent avoir que l'une des valeurs prédéfinies dans le questionnaire (par exemple, le sexe ne peut être codé que comme « 1 » pour les hommes ou « 2 » pour les femmes). Les variables chronologiques doivent contenir des dates valables, tandis que les variables numériques doivent se trouver à l'intérieur de valeurs maximum prescrites (par exemple entre 0 et 95 ans). La complétude peut également être vérifiée d'une autre façon lorsque les données provenant de deux ou plusieurs champs étroitement liés peuvent être comparées à des tableaux de références externes.

78. *Vérification des questions sautées*. Il s'agit en l'occurrence de déterminer si des codes établis concernant les questions à sauter ont été dûment suivis. Par exemple, une vérification simple tend à établir que des questions à poser uniquement à des enfants qui fréquentent l'école ne soient pas enregistrées pour un enfant qui a répondu « non » à la première question de savoir s'il va ou non à l'école. Selon son âge et son sexe, chaque membre de la famille est censé répondre (ou sauter) des questions spécifiques. Par exemple, les femmes de 15 à 49 ans peuvent être incluses dans la section du questionnaire concernant la fécondité, mais pas les hommes.

79. *Vérification de la cohérence.* Il s'agit de déterminer si les valeurs correspondant à une question cadrent avec celles provenant d'une autre question. La vérification est simple lorsque les deux valeurs concernent la même unité statistique, par exemple la date de naissance et l'âge d'un individu. Elle est plus complexe lorsqu'il faut comparer des informations provenant de deux ou plusieurs unités ou observations différentes. Par exemple, les parents doivent avoir au moins 15 ans de plus que leurs enfants.

80. *Vérification typographique.* Une erreur fréquente consiste à inverser des chiffres (par exemple inscrire « 14 » plutôt que « 41 ») dans une entrée numérique. S'il s'agit de l'âge, une telle erreur peut être détectée par une vérification de cohérence avec la situation conjugale ou les relations familiales. Par exemple, une indication d'erreur apparaîtra, dans le cas d'un adulte marié ou veuf de 41 ans dont l'âge est entré par erreur comme étant de 14 ans, lorsque l'âge sera comparé à la situation conjugale.

Cependant, la même erreur, s'agissant des dépenses mensuelles d'alimentation, risque aisément de ne pas être détectée, car 14 dollars ou 41 dollars peuvent être des montants plausibles. Pour éviter une telle situation, l'on demande généralement à deux opérateurs différents d'entrer deux fois les données provenant de chaque questionnaire.

Données manquantes

81. Lorsque commence le traitement, il y aura très certainement beaucoup de données manquantes. Il se peut que certains ménages aient déménagé ou aient refusé de répondre, qu'il n'ait pas été apporté de réponse à certaines des questions figurant dans le questionnaire, ou bien que certaines informations soient incompatibles avec les autres données figurant dans les réponses au questionnaire. Quelle que soit la raison, l'on se trouve en présence d'une rubrique manquante, vide ou incomplète.

82. Il importe d'établir une distinction entre les données manquantes — c'est-à-dire les données qui devraient être présentes mais dont la valeur correcte est inconnue — et les données néant. Par exemple, un questionnaire pourra être vide parce que le ménage a refusé de répondre à l'enquête, tandis qu'une partie d'un deuxième questionnaire pourra être vide parce que le ménage n'a rien semé dans ses champs. Dans le deuxième cas, la variable « superficie ensemencée » devra être indiquée comme étant néant. De telles entrées doivent être conservées dans le fichier pour analyse et tabulation.

83. La marche à suivre dans le cas de données véritablement manquantes dépend du type de données qui fait défaut. Un élément sélectionné de l'échantillon peut être totalement absent en raison du refus du ménage de participer à l'enquête ou de l'incapacité du déclarant de répondre à toutes les questions. En pareil cas, l'on dit qu'il y a eu « non-réponse unitaire ».

84. Si le déclarant ne peut répondre qu'à certaines des questions mais pas à d'autres, il y a « non-réponse ponctuelle/partielle » parce qu'il a été obtenu au sujet du ménage certaines informations, mais pas toutes.

85. Comme cela a été souligné à maintes reprises dans le présent guide, les données manquantes, quel qu'en soit le type, débouchent sur des estimations faussées. Pour une discussion détaillée du traitement réservé aux cas de non-réponse, y compris les méthodes d'ajustement, voir le chapitre 6.

86. En cas de non-réponse partielle, il peut être nécessaire, pour que les totaux correspondent, de remplacer les valeurs manquantes par des estimations raisonnables. C'est ce que l'on appelle l'imputation, comme on l'a vu au chapitre 6. Plusieurs méthodes peuvent être utilisées pour imputer des valeurs de remplacement. Certaines d'entre elles sont :

- *L'imputation de la valeur moyenne :* la valeur moyenne (de l'UPE ou de toutes les séries de données) est utilisée pour imputer la valeur manquante;
- *L'imputation séquentielle :* les valeurs manquantes sont empruntées à un dossier (donateur) semblable au dossier incomplet. Le dossier donateur doit avoir été dûment vérifié;
- *L'imputation statistique :* une relation (régression, ratio) est utilisée simultanément avec une autre variable tirée de données complètes pour imputer la valeur manquante.

87. Ces méthodes ne sont que quelques-unes de celles auxquelles il est possible d'avoir recours pour imputer des valeurs manquantes, mais il en existe plusieurs autres. L'efficacité de l'imputation dépendra évidemment de la mesure dans laquelle le module d'imputation capture la non-réponse.

Pour sélectionner les informations auxiliaires disponibles, il importe que la variable corresponde à celle qui doit être imputée (pour plus amples informations à ce sujet, voir Olsson (1990).

9.4.3.3. Entrée des donnés

88. L'entrée des données a pour but de transformer les informations figurant dans les questionnaires sur support papier en un produit intermédiaire (fichiers lisibles à la machine) qui devra ensuite être affiné par des programmes d'édition et des processus manuels pour pouvoir obtenir comme produit final des bases de données dites nettoyées. Au cours de la phase initiale de l'entrée des données, la priorité est accordée à la rapidité et à la nécessité de veiller à ce que l'information entrée dans les fichiers corresponde parfaitement à celle qui figure dans le questionnaire.

89. La méthode à utiliser pour entrer les données figurant dans les questionnaires doit être décidée à un stade aussi précoce que possible car elle aura un impact notable sur le déroulement du travail, la méthode de stockage des données, la conception du formulaire et la composition du personnel.

Entrée informatisée des données

90. L'entrée des données consiste à saisir des données codées sur un disque dur, une disquette ou un disque compact, par exemple. Les organismes d'enquête d'un grand nombre de pays en développement ont acquis une expérience considérable de cette méthode d'entrée des données. C'est la principale technique utilisée, surtout depuis le développement de l'ordinateur personnel et l'apparition de logiciels spécialisés.

91. *Application entrée des données.* Normalement, l'application entrée des données comporte trois modules. Le premier est celui dans lequel toutes les informations sont entrées. Le deuxième module, qui sert à vérifier les données entrées, est celui qui certifie que les informations entrées sont de bonne qualité et qui vérifie le travail des opérateurs chargés de l'entrée des données. Le troisième module sert à corriger les informations entrées s'il est nécessaire de rectifier des erreurs qui n'ont pas été détectées lors des processus d'entrée ou de validation des données.

92. L'application entrée des données comporte habituellement un menu principal dans lequel le préposé peut sélectionner l'opération à effectuer: entrée, vérification et correction des données. Avant de travailler dans le menu principal, l'usager doit certifier, au moyen d'un mot de passe, qu'il est autorisé à pénétrer dans l'application. Si le nom de l'utilisateur ou le mot de passe est incorrect, l'application doit se fermer automatiquement. Le nom de tous les usagers et leurs mots de passe sont stockés dans un tableau d'usagers, et les mots de passe sont chiffrés. Lorsqu'un usager entre dans le système avec un mot de passe valable, cela est consigné dans le tableau.

93. *Module d'entrée des données.* Le module d'entrée des données est l'interface entre le questionnaire et le fichier ou la base de données. Le système doit être très simple pour pouvoir être utilisé facilement par le préposé à l'entrée des données. Certaines des règles à observer sont notamment les suivantes :

- L'écran d'entrée des données doit ressembler d'aussi près que possible aux pages correspondantes du questionnaire. L'opérateur doit pouvoir retrouver très rapidement, à partir du questionnaire, le champ correspondant figurant sur l'écran.
- La rapidité est très importante au stade de l'entrée des données. L'opérateur ne doit pas attendre que le système évalue chacune des valeurs entrées. Le processus d'évaluation doit

par conséquent être extrêmement rapide, ce qui signifie que le système ne doit pas avoir plus de contacts avec le serveur que nécessaire, et que les valeurs ne seront sauvegardées dans la base de données que lorsque toutes les valeurs correspondantes au ménage correspondant auront été entrées. L'inconvénient est que les informations venant d'être entrées disparaîtront si, pour une raison quelconque, l'application est fermée. Toutefois, cet inconvénient est plus que compensé par l'avantage d'une rapidité relativement élevée.

- Chacune des valeurs consignées dans le questionnaire doit être assortie d'un code numérique pour pouvoir utiliser le clavier numérique et ainsi procéder plus rapidement.
- Le module d'entrée des données doit avoir un contrôle de validité variable qui signale immédiatement à l'opérateur l'entrée d'une valeur non valable. Les contrôles de validité doivent également englober les valeurs connexes; par exemple, si le « sexe » a la valeur « 1 » (hommes), il ne doit pas être possible d'entrer des informations concernant la fécondité.
- Le programme d'entrée des données doit évidemment signaler comme erreur toute situation qui constitue une impossibilité logique ou naturelle (par exemple une fille qui serait plus âgée que sa mère) ou qui sont très peu probables (comme un écart de moins de 15 ans entre l'âge de la mère et celui de la fille).
- Il importe d'enregistrer le nombre de caractères entrés et la durée du processus d'entrée des données afin de pouvoir, par exemple, prédire le temps que prendra une opération future.

94. *Module de vérification des données*. Un système de vérification a pour objet de fournir des informations sur la qualité des données entrées et la proportion des erreurs commises par chaque opérateur. L'écran de ce module se présente exactement comme celui du module d'entrée des données, sans aucune différence visible. La principale différence est que les modules enregistrent non seulement le nombre de caractères entrés mais aussi le nombre d'erreurs. Les options disponibles en ce qui concerne le type de vérification sont notamment la *vérification totale*, qui porte sur toutes les *zones d'énumération* et sur tous les questionnaires remplis dans la *zone d'énumération* considérée, ou la *vérification par sondage,* qui ne porte que sur quelques-unes des *zones d'énumération* ou quelques-uns des questionnaires.

95. *Module de correction des données.* Ce module est utilisé principalement pour corriger les informations qui, pour une raison ou pour une autre, n'ont pas pu être complétées dans le module d'entrée des données. L'on peut, avec ce module, ajouter, supprimer et mettre à jour des informations, qu'elles concernent l'ensemble d'un ménage ou une seule valeur.

96. *Application administration.* L'application administration est l'outil au moyen duquel les superviseurs peuvent introduire des modifications dans la base de données. Cet outil sert principalement à corriger le fichier-maître par lots et à suivre le travail des usagers. Il faut :

- Que les superviseurs aient un contrôle complet du fichier-maître à partir de cette application et qu'ils puissent ajouter, supprimer et mettre à jour les informations qu'il contient.
- Que l'on puisse ajouter et supprimer des usagers et obtenir une liste complète de tous les usagers et que l'on puisse également vérifier le statut actuel de tous les usagers ou d'un seul uniquement.
- Qu'il soit possible de contrôler et d'imprimer des statistiques concernant l'entrée des données sur des périodes différentes.

- Qu'il soit possible de contrôler et d'imprimer le taux d'erreurs d'un seul usager et le taux moyen pour tous les usagers.
- Qu'il soit possible de revenir à une *zone d'énumération* pour entrer ou vérifier des données.
- Que cette application permette aux superviseurs d'obtenir toutes les informations dont ils ont besoin pour gérer leur travail.

Svensson (1996) expose en détail les divers aspects des systèmes automatisés d'entrée des données.

97. *Plates-formes utilisées pour l'entrée automatisée des données*. Il existe dans le commerce un grand nombre de plates-formes qui peuvent être utilisées pour l'entrée et l'édition des données. Par exemple, le Census and Survey Processing System et son prédécesseur, l'IMPS, se sont avérés être des systèmes efficaces pour l'entrée et l'édition de données dans le contexte d'enquêtes nationales complexes réalisées dans de nombreux pays en développement. Il s'agit également de systèmes que l'on peut se procurer et utiliser facilement (Munoz, 2003).

Numérisation

98. Il y a quelques années encore, l'entrée des données se faisait habituellement sur clavier et l'on ne trouvait pas dans le commerce nombre de systèmes concurrents. Aujourd'hui, les choses ont changé du tout au tout et les systèmes d'entrée de données les plus répandus sont tous fondés sur des techniques de numérisation, ayant chacune ses propres avantages et ses propres inconvénients. Les méthodes les plus communément utilisées sont la reconnaissance optique de caractères (OCR), système qui permet de lire des caractères imprimés à la machine; la reconnaissance intelligente de caractères (ICR), qui reconnaît des caractères manuscrits; la reconnaissance optique de marques (OMR), système qui peut lire des marques portées à l'encre ou au crayon dans des positions prédéterminées, habituellement des cases; et des codes à barres (BCR), système qui lit les données codées figurant à l'intérieur des barres.

99. Selon Lundell (2003), les deux systèmes qui se prêtent le mieux aux enquêtes statistiques et aux recensements sont les systèmes de reconnaissance intelligente de caractères et les systèmes de reconnaissance optique des marques. Un pays comptant une population nombreuse tendrait à préférer ce dernier système, tandis que le système de reconnaissance intelligente de caractères est plus indiqué dans le cas de questionnaires complexes. Le système de reconnaissance optique de marques impose certaines contraintes en ce qui concerne la conception du formulaire mais permet un traitement rapide et n'exige que du personnel relativement moins qualifié. Le système de reconnaissance intelligente de caractères peut être utilisé quelle que soit la conception du formulaire mais le traitement est plus exigeant pour ce qui est des capacités informatiques et des aptitudes du personnel. Les codes à barres ne sont généralement utilisés que pour imprimer et rechercher des informations sur l'identité, par exemple, les numéros de formulaire, étant donné que le code à barres contient un système intégré de vérification afin de minimiser les erreurs.

100. Les questionnaires sont numérisés à des vitesses comprises entre 40 et 90 feuilles par minute en duplex. La rapidité est le principal avantage de la numérisation par rapport aux formes traditionnelles d'entrée des données sur clavier. Le logiciel de numérisation est utilisé pour identifier les pages du questionnaire et en évaluer le contenu au moyen de systèmes de reconnaissance intelligente de caractères ou de reconnaissance optique de marques. Les informations douteuses ou les informations à coder sont envoyées au vérificateur, qui revoit les informations peu lisibles et code les réponses ouvertes en se référant aux tableaux électroniques intégrés au modèle de numérisation. Ces véri-

fications peuvent être accomplies de manières extrêmement diverses, selon la structure du logiciel de numérisation. Les variables critiques peuvent être prévues en tout ou en partie pour maximiser l'exactitude des réponses entrées dans le fichier de données.

101. Il est généralement admis que l'utilisation d'un système de numérisation peut accroître dans des proportions allant jusqu'à 70 % l'efficacité du processus de capture des données (Edwin, 2003). Nombre des problèmes qui peuvent surgir peuvent être éliminés si le processus de numérisation est bien organisé du point de vue technique. Par exemple, les problèmes qui se posent lorsque des pages manquent ou ne se suivent pas peuvent être résolus en pré-imprimant des codes à barres sur les questionnaires et en utilisant ces codes pour relier les différentes pages d'un même questionnaire. Si le matériel et les logiciels sont convenablement entretenus et gérés, le coût final de la numérisation (y compris l'achat du matériel et des logiciels) peut être bien inférieur à celui d'une opération d'entrée manuelle des données.

102 Les systèmes de numérisation ont généralement été très peu utilisés pour la réalisation d'enquêtes sur les ménages, surtout en Afrique subsaharienne. Ces systèmes ont cependant été très largement utilisés lors de la série de recensements de la population et du logement de 2000, ce qui marque peut-être un tournant vers leur adoption généralisée. L'Afrique du Sud, le Kenya, la Namibie, la République-Unie de Tanzanie et la Zambie, par exemple, ont utilisé des systèmes de numérisation lors de leurs derniers recensements. Ces systèmes ont récemment été utilisés aussi pour les enquêtes sur les indicateurs du bien-être réalisées sous l'égide de la Banque mondiale. Des pays comme l'Afrique du Sud et la Namibie les ont également adoptés pour leurs programmes d'enquêtes sur les ménages.

9.4.3.4. Structure des fichiers et organisation des séries de données

Stockage des données

103. Dans le cas d'enquêtes sur les ménages, pour lesquelles il est habituellement rassembler des informations aux niveaux aussi bien des ménages que des individus, le mieux, pour utiliser efficacement la mémoire disponible, consiste à utiliser un fichier séquentiel ou un fichier de longueur variable car le nombre de personnes qui composent les ménages varie. Un fichier « plat », qui occuperait inutilement de la place dans la mémoire, ne serait indiqué que si toutes les questions se référaient au ménage en tant qu'unité statistique mais, comme on l'a vu, tel n'est pas le cas. Certaines des questions se rapportent à des unités statistiques subordonnées et apparaissent en nombres variables au sein de chaque ménage, comme personnes, cultures, articles de consommation, etc. Stocker les données concernant l'âge et le sexe de chaque membre du ménage comme variables différentes au niveau des ménages serait un gaspillage étant donné que les variables requises seraient définies par la taille du ménage le plus nombreux plutôt que par la taille du ménage moyen.

104. Un fichier de longueur variable serait normalement utilisé pour l'entrée des données d'enquêtes sur les ménages. Comme la taille et la composition des ménages varient, les fichiers utilisés pour l'entrée des données doivent être de longueur variable. Les données seront de longueur et de format fixes, mais chaque fichier contiendra des données de types différents. Chaque fichier sera essentiellement une image informatisée des questionnaires remplis. Chaque ligne ou case du questionnaire constituera une entrée. Chaque entrée commencera par une série d'indications reliant l'entrée au ménage ou à l'unité d'observation dont il s'agit. Cette méthode utilise efficacement l'espace disponible mais est peu commode lors du traitement ultérieur, lorsqu'il devient indispensable d'opérer des vérifications croisées de données se trouvant dans des fichiers différents.

105. Le système CSPro, par exemple, utilise une structure de fichier qui fait face efficacement aux complexités inhérentes à la présence d'un grand nombre d'unités statistiques différentes tout en minimisant la mémoire nécessaire et en assurant une interface efficace avec les logiciels statistiques au stade de l'analyse.

106. La structure des données maintient une correspondance directe entre chaque unité statistique observée et les entrées figurant dans les fichiers en utilisant un type d'entrée différente pour chaque type d'unité statistique. Par exemple, pour gérer les données figurant sur la liste de ménages, il serait défini un type déterminé d'entrée pour les variables figurant sur la liste et les données correspondant à chaque individu seraient stockées dans une entrée distincte du même type. De même, dans le module de la consommation alimentaire, une entrée type correspondrait aux types d'aliments et les données correspondant à chaque aliment seraient stockées dans les entrées distinctes du même type.

107. Le nombre d'entrées de chaque type peut varier, ce qui économise la mémoire requise étant donné que les fichiers n'ont pas à devoir systématiquement accepter les entrées les plus volumineuses possibles.

108 Après les données d'identification, les informations proprement dites concernant chaque unité sont entrées dans des champs de longueur fixe dans le même ordre que celui des questions posées. Toutes les données sont stockées sous le format standard ASCII (American Standard Code for Information Interchange).

109. Les ouvrages de Munoz (2003) et de la Banque mondiale (1991) contiennent des informations plus détaillées touchant la gestion des fichiers dans le contexte des enquêtes sur les ménages.

Restructuration des séries de données pour des opérations ultérieures

110. Pour faciliter l'analyse, la base de données doit contenir toutes les informations requises concernant la procédure d'échantillonnage, les étiquettes pour les strates d'échantillonnage, les unités primaires d'échantillonnage, les unités secondaires d'échantillonnage, etc., et les pondérations pour chaque unité d'échantillonnage. Ces informations seront nécessaires pour estimer les statistiques requises ainsi que les erreurs d'échantillonnage de ces estimations.

111. Après l'entrée des données, il faut fréquemment restructurer les séries de données et générer de nouveaux fichiers et coder à nouveau certains des champs de données existants de manière à définir de nouvelles variables d'une manière plus commode pour la tabulation et l'analyse, par exemple pour pouvoir réaliser certaines opérations sur les données, par exemple établir des estimations.

112. Il se peut que le fichier initial contenant toutes les données provenant de l'enquête contienne en fait des informations sur des unités appartenant à des populations différentes (Rosen, 1991). Dans le cas d'une enquête sur le budget des ménages, par exemple, le même fichier initial pourra comprendre les données concernant les ménages faisant partie de l'échantillon ainsi que les personnes interrogées. Pour estimer les caractéristiques statistiques de la population de ménages et de la population d'individus, il faut avoir un dossier contenant une entrée pour chaque ménage et un dossier contenant une entrée pour chaque individu respectivement. Les séries ou fichiers de données basés sur les ménages comme unités ou objets sont utilisés pour produire des statistiques (tableaux, etc.) sur les ménages privés. Les séries ou fichiers de données basés sur des individus en tant qu'unités (objets) sont utilisés pour produire des statistiques (tableaux) sur les individus faisant partie des ménages privés.

113. Comme le montre clairement ce qui précède, il y a habituellement deux principaux types de fichiers: les fichiers de ménages et les fichiers d'individus. Le plus souvent, les fichiers se composent

de fichiers de ménages en ce sens qu'ils contiennent des données touchant les variables liées à l'unité d'observation ou au « ménage objet ». Certains sont des fichiers d'individus en ce sens qu'ils contiennent des données se rapportant aux variables liées à l'unité d'observation ou « individu objet ». Les fichiers de données complets et définitifs (séries de données) contiendront des informations sur tous les ménages déclarants et sur les individus appartenant à chacune des unités primaires d'échantillonnage visées.

114. Le tableau 9.2 illustre comment le volumineux fichier utilisé pour l'Enquête démographique intercensitaire réalisée au Zimbabwe en 1987 a été réorganisé pour faciliter un traitement plus poussé des données. Le deuxième exemple du tableau 9.3 présente les fichiers habituellement utilisés pour une enquête sur le budget des ménages. Ces exemples sont fondés sur les données analysées par Lagerlof (1988) et Rosen (1991).

Tableau 9. 2
Fichiers de ménages et fichiers d'individus utilisés pour l'Enquête démographique intercensitaire réalisée au Zimbabwe en 1987

Fichier	Type	Contenu
MÉNAGE	Fichier de ménages	Identification des ménages (région, province, district, etc.) Réponses à toutes les questions concernant le ménage Variables dérivées, comme taille du ménage (tirée du fichier concernant ses membres), etc.
PERSONNE	Fichier d'individus	Identification du ménage (HID) plus identification de la personne (PID) Caractéristiques démographiques: ÂGE, SEXE, MARSTAT (situation conjugale), USMEM (membre habituel du ménage), RELTH (relation avec le chef de ménage)
DÉFUNT	Fichier d'individus	Identification du ménage (HID) plus identification du défunt (DID) Informations sur le défunt qui faisait habituellement partie du ménage : SEXE, ÂGED (âge du défunt)
FEMME ≥ 12 ans	Fichier d'individus	HID, PID Informations sur toute femme de plus de 12 ans faisant partie du ménage

Tableau 9.3
Fichiers habituellement utilisés pour une enquête sur le budget des ménages

Fichier	Type	Contenu
MÉNAGE	Fichier de ménages	Identification des ménages (région, province, district, etc.) Réponses à toutes les questions concernant le ménage Variables dérivées, comme taille du ménage (tirée du fichier concernant ses membres), etc.
MEMBRES	Fichier d'individus	Identification du ménage plus identification du membre du ménage Caractéristiques démographiques: âge, sexe, situation conjugale, degré d'instruction, etc. Informations concernant les principales activités : situation au regard de l'emploi, profession, etc.
REVENU	Fichier d'individus	Identification du ménage plus identification de la personne plus identification de la source de revenu

Fichier	Type	Contenu
ALIMENTATION	Fichier d'individus	Identification du ménage plus identification du type d'aliment Dépenses d'alimentation :
AUTRES ARTICLES NON DURABLES	Fichier de ménages	Identification du ménage plus identification de l'article non durable Dépenses en articles non durables :
ARTICLES DURABLES	Fichier de ménages	Identification du ménage plus identification de l'article durable Dépenses en articles durables :
AGRICULTURE	Fichier d'individus	Identification du ménage plus identification du type d'agriculture Dépenses consacrées à l'agriculture :
ÉQUIPEMENT AGRICOLE	Fichier d'individus	Identification du ménage plus identification du type d'équipement agricole Dépenses d'équipement agricole :

115. Pour la tabulation, la plupart des logiciels statistiques exigent un fichier « plat ». La plupart des logiciels généralement disponibles exigent des données se présentant sous ce format. Dans un fichier plat, toutes les entrées ont la même série de variables ou de champs et sont de même longueur. Un fichier est considéré comme « plat » lorsqu'il existe pour chaque déclarant exactement la même série de champs de données. Les champs de données sont agencés de la même façon pour chaque entrée, le fichier comprend un nombre fixe d'entrées agencées de la même façon. Le tableau 9.4 illustre le format du fichier plat de ménages utilisé pour l'Enquête démographique intercensitaire réalisée au Zimbabwe en 1987.

116. Le fichier de ménage contient une entrée pour chaque ménage observé, chaque entrée contenant des informations sur :

- L'identification du ménage;
- Les paramètres de conception d'échantillonnage;
- Les valeurs observées des variables concernant les ménages;
- Les variables de pondération.

Tableau 9.4
Format de fichier plat de ménages utilisé pour l'Enquête démographique intercensitaire réalisée au Zimbabwe en 1987.

Identification				Paramètres de la conception d'échantillonnage							Valeurs variables			Variable de pondération
Strate	Subdivision	ZE	Hh	S_h	a_h	R_h	b_{hr}	S_{hi}	M_{hi}	m_{hi}	x	y	z	w
h	r	i	h								x_{hrij}	y_{hrij}	z_{hrij}	w_{hrij}

Identification du ménage : la combinaison *hrij* signifie que le ménage *j* appartient à la *zone d'énumération* (*ZE*) *i* de la subdivision *r* de la strate *h.*

Paramètres d'échantillonnage : dans cet exemple spécifique, les paramètres d'échantillonnage étaient les suivants :

- S_h = nombre de ménages faisant partie de la strate d'échantillonnage en 1982
- a_h = taille de la *ZE* sélectionnée faisant partie de la strate d'échantillonnage
- R_h = nombre de subdivisions représentées dans l'échantillon tiré de la strate d'échantillonnage
- b_{hr} = nombre de *ZE* sélectionnées dans la subdivision
- S_{hi} = nombre de ménages faisant partie de la *ZE* en 1982
- M_{hi} = nombre de ménages faisant partie de la *ZE* en 1987
- m_{hi} = taille du ménage faisant partie de l'échantillon de la *ZE.*

Valeurs variables observées: dénotent les variables des ménages.

Variables de pondération: dénote la variable de pondération pour le ménage**.**

117. Les fichiers d'individus sont organisés comme les fichiers de ménages présentés ci-dessus, sous réserve d'une différence mineure, à savoir que l'identification concernera la personne (PID) et que l'index (*k*) se rapportera à la personne visée, tandis que les « variables » se réfèreront aux variables des individus.

118 Au stade de leur diffusion, il suffit d'organiser les données en fichiers plats distincts (un pour chaque type d'entrées) car le format de longueur fixe de la structure originelle vaut aussi pour le transfert des données à des systèmes standard de gestion des données (DBMS) pour d'autres opérations de traitement ou à des logiciels standard à des fins de tabulation et d'analyse. Il est très facile de transférer les données à des systèmes de gestion de base de données car la structure originelle peut être traduite presque directement sous le format standard DBF que tous les systèmes acceptent pour l'entrée de tableaux (dans ce cas particulier, les codes d'identification des entrées constituent les liens naturels entre les tableaux) (Munoz, 2003).

9.4.3.5. Procédures des estimations et calcul des pondérations

119. Le chapitre 6 a donné une description détaillée de la raison d'être et de la méthode de calcul des pondérations des données d'enquêtes sur les ménages (voir également les ouvrages de Rosen cités à la fin du présent chapitre). Un algorithme de calcul qui, à partir des valeurs observées, permet d'obtenir des estimations des caractéristiques statistiques, est l'une des procédures d'estimation ponctuelle qui peuvent être suivies. Pour commencer, il est calculé une pondération pour chaque objet déclarant. Il est ensuite calculé des estimations des « totaux » en ajoutant les valeurs d'observation pondérées (valeurs observées multipliées par la pondération correspondante).

120. Munoz (2003) donne une bonne description de l'utilisation qui peut être faite du système Microsoft Excel pour appliquer la procédure d'échantillonnage tout au long de ses principales phases: organisation du cadre à utiliser pendant la première phase; sélection des unités primaires d'échantillonnage sur la base d'une probabilité proportionnelle à la taille; et calcul des probabilités de sélection et des pondérations d'échantillonnage correspondantes.

121. La construction proprement dite des estimations pondérées est simple. L'on commence par la série de données originelle et l'on créé ensuite une nouvelle série de données en multipliant chaque observation par le nombre de fois spécifié par sa pondération avant d'utiliser les formules standard pour calculer le paramètre considéré sur la base de la série de données pondérées.

122. Il y a lieu de noter toutefois que, pour être exactes, les pondérations doivent comporter trois éléments (Yansaneh, 2003), y compris les divers ajustements requis (voir aussi le chapitre 6). Les pondérations de base tiennent compte de la variation des probabilités de sélection des différents groupes

de ménages, comme stipulé dans la conception initiale de l'enquête. Le deuxième ajustement a pour but de tenir compte de la variation des taux de non-réponse entre les différents domaines ou sous-groupes. Enfin, il peut être nécessaire dans certains cas de procéder à des ajustements post-stratification pour que les données provenant de l'enquête soient conformes aux répartitions tirées d'une source indépendante, comme le dernier recensement de la population.

123. Une autre complication du processus d'estimation découle de la demande croissante de statistiques au niveau des domaines. Comme on l'a dit au chapitre 3, un domaine est un sous-groupe pour lequel l'on souhaite obtenir des estimations distinctes. Habituellement, ces sous-groupes peuvent être spécifiés au stade de la conception de l'échantillon, mais ils peuvent aussi être définis sur la base des données dérivées. Un domaine peut également être une strate, une combinaison de strates, des régions administratives (province, district, milieu rural/milieu urbain) et peut également être défini en termes de caractéristiques démographiques ou socioéconomiques (par exemple âge, sexe, groupe ethnique, pauvreté, etc.). L'on essaiera ci-après de décrire comment des séries de données peuvent être construites pour faciliter l'élaboration d'estimations au niveau de domaines.

124. L'on commencera par visualiser un fichier de données d'observations (par exemple le fichier de ménages), comme indiqué ci-dessus dans le contexte de l'Enquête démographique intercensitaire réalisée au Zimbabwe. Ce fichier comporte une entrée pour chaque ménage sélectionné. À la fin de l'enquête, le fichier contiendra pour chaque ménage les informations suivantes :

a) Code d'identification des ménages;

b) Paramètres d'échantillonnage;

c) Valeurs des variables X, Y et Z à étudier;

d) Valeur de la pondération d'estimation du ménage;

e) Appartenance ou non-appartenance du ménage à la catégorie C;

f) Appartenance ou non-appartenance du ménage au domaine G.

125. Ces informations (hormis les paramètres d'échantillonnage) sont dénotées comme suit :

- HID = étiquette d'identification des ménages sélectionnés. Dans un souci de simplicité, nous suivrons l'ordre numérique, 1, 2, ..., n. Par conséquent, n dénote la taille de l'ensemble de l'échantillon;
- x, y et z sont les valeurs observées des variables X, Y et Z du ménage;
- c = 1 si le ménage appartient à la catégorie C, autrement il est égal à 0;
- g = 1 si le ménage appartient au domaine G, autrement il est égal à 0;
- w = la pondération de l'estimation concernant le ménage.

126. Les valeurs des variables indicatives C et G sont habituellement dérivées de celles d'autres variables et ne sont pas observées directement. Par exemple, la catégorie C pourrait être la catégorie « au-dessous du seuil de pauvreté ». Il n'est pas demandé au ménage s'il appartient à cette catégorie ou non. La classification est tirée, par exemple, des données concernant le revenu du ménage et d'un seuil de pauvreté prédéterminé. De même, il faut souvent dériver des données d'autres variables pour déterminer si un ménage appartient à un domaine d'étude spécifique G ou non (par exemple le domaine G peut être composé de ménages ayant trois enfants ou plus). Au stade de l'estimation, les valeurs de ces indicateurs doivent être disponibles dans le fichier d'observation.

127. Lorsque toutes les données sont disponibles dans le fichier d'observations, celui-ci apparaîtra comme indiqué au tableau 9.5 ci-dessous, sauf que les paramètres d'échantillonnage ne sont pas inclus.

Tableau 9.5
Fichier d'observations contenant les données finales concernant les variables de l'enquête sur les ménages

HID	X	Y	Z	C	G	W
1	x_1	y_1	z_1	c_1	G_1	w_1
2	x_2	y_2	z_2	c_2	G_2	w_2
3	x_3	y_3	z_3	C_3	G_3	w_3
.	.	.		.	.	.
w	.	.		.	.	.
N	x_n	y_n		c_n	g_n	w_n

128. La discussion ci-dessus a porté uniquement sur l'estimation des caractéristiques statistiques de la population de ménages. Les estimations des caractéristiques statistiques pour la population de personnes sont effectuées de la même façon. Généralement, la pondération de l'estimation concernant une personne est identique à celle du ménage dans laquelle la personne en question appartient. Si tous les membres d'un ménage type sont énumérés dans le questionnaire, une personne déterminée n'est incluse dans l'échantillon de personnes que si et seulement si le ménage auquel elle appartient fait partie de l'échantillon de ménages. De ce fait, la probabilité d'inclusion d'une personne est identique à la probabilité d'inclusion du ménage auquel elle appartient. Il y a lieu de noter toutefois que ce qui précède n'est pas vrai lorsque l'on procède à un sous-échantillonnage parmi les ménages. Par exemple, selon certaines conceptions, il se peut que la procédure envisage de ne sélectionner qu'un adulte par ménage ou un homme et une femme. En pareil cas, la pondération de l'individu ou des individus sélectionnés est calculée de façon indépendante et n'est pas égale à celle du ménage.

129. Pour des raisons de complétude, la procédure de l'estimation suivie dans le contexte d'une enquête sur les ménages doit prévoir l'établissement d'estimations des erreurs d'échantillonnage (ou des erreurs types), surtout pour les statistiques les plus importantes qui sont publiées. Tout le chapitre 7 du guide est consacrée à cette question.

9.4.3.6. Tabulation, séries de données utilisées pour la tabulation et bases de données

130. Les principaux produits d'une enquête statistique sont au nombre de trois (Sundgren, 1995) :

- *Macrodonnées :* « statistique » représentant des estimations de certaines caractéristiques statistiques; la production de ces données est le principal objectif de l'enquête réalisée;
- *Microdonnées :* « observations d'objets individuels » à la base des macrodonnées produites par l'enquête. Ces données sont essentielles pour pouvoir utiliser et interpréter ensuite les résultats de l'enquête;
- *Métadonnées :* « données décrivant la signification, l'exactitude, la disponibilité et les autres principaux attributs des micro et des macrodonnées sous-jacentes »; les métadonnées sont essentielles pour pouvoir identifier correctement et rechercher les données statistiques concernant un problème spécifique ainsi que pour interpréter correctement et (ré)utiliser les données statistiques.

Il serait utile aussi de considérer la conception de tableaux multidimensionnels (cubes) afin de pouvoir avoir accès plus facilement aux résultats de l'enquête et s'y référer, par exemple par le biais de sites Web.

131. À terme, le programme d'enquêtes sur les ménages devra déboucher sur une situation telle que l'archivage des données soit fondé sur une combinaison de micro et de macrodonnées. À cette fin, il faut décrire en détail la structure des informations collectées au moyen d'enquêtes multiples.

132. Le stockage des données doit être considéré comme comportant trois phases (Lundell, 2003) :

- *Stockage :* lors de l'entrée des données, celles-ci doivent être stockées de la manière qui correspond le mieux aux méthodes d'entrée et de nettoyage des données utilisées, comme indiqué ci-dessus;
- *Entreposage :* lorsque les données ont été entrées et nettoyées, il faut les ajouter à un entrepôt dont la structure soit adaptée aux outils et aux méthodes d'analyse et de diffusion des données;
- *Archivage :* les données doivent être archivées conformément aux normes établies pour qu'elles puissent être aisément recherchées à l'avenir.

133. L'on peut créer de différentes façons un entrepôt contenant des données nettoyées en ayant recours à l'une des méthodes suivantes (voir l'appendice au présent chapitre pour plus amples informations sur ces logiciels) :

- Fichiers plats
- Bases de données relationnelles (par exemple le serveur Structured Query Language (SQL) de Microsoft)
- Des logiciels statistiques [par exemple le système d'analyse statistique (SAS) et le Statistical Package for the Social Sciences (SPSS)]

134. Pour l'archivage à long terme des données finales, le mieux est de sauvegarder les données dans les fichiers plats sous format ASCII avec une description jointe des entrées. La plupart des systèmes de gestion de bases de données et des logiciels statistiques peuvent exporter des données vers ces fichiers sans guère de difficultés et peuvent aussi en importer aisément.

Appendice

Logiciels pouvant être utilisés aux différentes étapes du traitement des données d'enquête

Type d'opération	Logiciel
Système de gestion de bases de données	Serveur Structured Query Language (SQL) de Microsoft 2000, édition standard Microsoft Access Statistical Analysis System (SAS)
Entrée et édition des données	Visual Basic Microsoft Access Integrated Microcomputer Processing System (IMPS) Census and Survey Processing System (CSPro)
Recherche des données	Statistical Analysis System (SAS) Statistical Package for the Social Sciences (SPSS) Microsoft Access Microsoft Excel
Tabulation, analyse et présentation	Microsoft Word Microsoft Excel Statistical Analysis System (SAS) Statistical Package for Social Sciences (SPSS)
Estimation des variance	CENVAR: élément calcul des variances du système IMPS. Computation and Listing of Useful STatistics on Errors of Sampling (CLUSTERS) Integrated System for Survey Analysis (ISSA) Survey Data Analysis (SUDAAN) Statistical Analysis System (SAS) Statistical Package for the Social Sciences (SPSS) Cluster Analysis and Regression Package (PC-CARP)

Microsoft Office

Microsoft Office, mis au point par Microsoft Corporation, est un jeu de logiciels contenant différents programmes, dont :

- Microsoft Office Access, qui est le programme de gestion de la base de données Office, particulièrement facile à utiliser et comprend des possibilités accrues d'importer et d'exporter des fichiers de données et de travailler sur des fichiers Extensible Markup Language (XML).
- Microsoft Office Excel, programme de tabulation comportant un appui pour les fichiers XML qui offre de nouvelles caractéristiques permettant d'analyser et de partager plus facilement des informations.
- Microsoft Office Word, qui est le programme de traitement de texte.

- Microsoft SQL Server 2000 est le serveur utilisé pour la gestion de l'ensemble des activités et des ressources d'une organisation.
- Microsoft Office Outlook, gestionnaire personnel de l'information et programme de communication Office qui offre une plate-forme unifiée pour la gestion du courriel, des calendriers, etc.

Site Web : http://www.microsoft.com/office/system/overview.mspx#EDAA

Visual Basic

Microsoft a lancé Visual Basic en 1987. Visual Basic n'est pas seulement un langage de programmation mais aussi un environnement de développement graphique complet. Cet environnement permet à l'usager n'ayant guère d'expérience de la programmation de développer **rapidement** les applications Microsoft Windows utiles pouvant utiliser des objets OLE (Object Linking and Embedding), comme des tableaux Excel. Visual Basic offre également la possibilité d'élaborer des programmes qui peuvent être utilisés comme application principale d'un système de gestion de base de données en constituant l'interface qui collecte les entrées de l'usager et affiche le produit formaté sous une forme plus attrayante et plus utile que celle que peuvent offrir nombre de versions SQL.

Le principal attrait de Visual Basic est la facilité avec laquelle l'usager peut créer des programmes graphiques bien conçus sans guère de codage du programmeur. Le principal objet, dans Visual Basic, est appelé un **formulaire,** ce qui facilite la préparation d'écrans d'entrée de données.

Site Web : http://www.engin.umd.umich.edu/CIS/course.des/cis400/vbasic/vbasic.html.

CENVAR

CENVAR est la composante de calcul des variances de l'Integrated Microcomputer Processing System (IMPS), série de logiciels utilisés pour l'entrée, l'édition, la tabulation, l'estimation, l'analyse et la diffusion des données provenant de recensements et d'enquêtes. Le système IMPS a été mis au point par le Census Bureau des États-Unis.

Site Web : http://www.census.gov/ipc/www/imps/

PC CARP

CENVAR est basé sur le logiciel Cluster Analysis and Regression Package for Personal Computers (PC CARP) initialement mis au point par l'Université de l'État de l'Iowa. PC CARP utilise la procédure de linéarisation pour le calcul des variances.

Site Web : http://www.census.gov/ipc/www/imps/

Census and Survey Processing System (CSPro)

CSPro (Census and Survey Processing System) est un logiciel du domaine public qui sert à l'entrée, l'édition, la tabulation et la cartographie des données provenant de recensements et d'enquêtes. Le système CSPro est le résultat d'un effort mené conjointement par les concepteurs des systèmes IMPS et ISSA : le Census Bureau des États-Unis, Macro International et Serpro, S.A. Les travaux de mise au point de ce logiciel ont été financés par le Bureau pour la population de l'Agency for International

Development des États-Unis. Le système CSPro a été conçu de manière à remplacer peu à peu les systèmes aussi bien IMPS que ISSA.

Site Web : http://www.census.gov/ipc/www/imps/

Computation and Listing of useful Statistics on Errors of Sampling (CLUSTERS)

Le système Computation and Listing of Useful Statistics on Errors of Sampling (CLUSTERS) a été mis au point initialement pour calculer les erreurs d'échantillonnage dans le contexte du programme mondial d'enquêtes sur la fécondité. Il utilise la méthode de linéarisation de Taylor pour le calcul des erreurs d'échantillonnage. Ce système a également été utilisé pour calculer les erreurs d'échantillonnage de différentes enquêtes sur les ménages, surtout celles visées dans le cadre des programmes d'enquêtes démographiques et sanitaires, dans de nombreux pays en développement (voir Verma, 1982).

Integrated System for Survey Analysis (ISSA)

L'Integrated System for Survey Analysis (ISSA) a été mis au point par Macro International Inc. expressément pour le programme d'enquêtes démographiques et sanitaires. Ce système a été utilisé pour les aspects du traitement, de l'entrée, de l'édition et de la tabulation des données. Il comporte également un module qui permet de calculer les erreurs d'échantillonnage de mesures démographiques complexes comme les taux de fécondité et de mortalité selon la méthode « jackknife » [voir Macro International Inc. (1996)].

Statistical Analysis System (SAS)

Le Statistical Analysis System (SAS), mis au point par SAS, Inc. en 1966, est un logiciel d'analyse des données, de gestion des fichiers et de calcul des erreurs d'échantillonnage [voir An et Watts (2001) pour un exposé de certaines des dernières caractéristiques du système SAS].

Statistical Package for the Social Sciences (SPSS)

Le Statistical Package for the Social Sciences (SPSS), mis au point par SPSS, Inc., est un logiciel d'analyse des données, de gestion des fichiers, etc. [voir SPSS, Inc. (1988) pour une présentation de certaines des dernières caractéristiques du système].

Survey Data Analysis

Le système Survey Data Analysis (SUDAAN), mis au point par le Research Triangle Institute (Research Triangle Part, Caroline du Nord), est un logiciel complet d'échantillonnage et d'analyse des données connexes qui se prête particulièrement à des analyses aussi bien descriptives que de modélisation. [L'on trouvera de plus amples détails dans Shah, Barnwell et Bieler (1996.)]

Références et autres lectures

An, A. et D. Watts (2001). *New SAS Procedures for Analysis of Sample Survey Data.* SUGI paper, n° 23. Cary, Caroline du Nord, SAS Institute, Inc.

Arnic *et al.* (2003). « Metadata production systems within Europe: the case of the statistical system of Slovenia », document présenté lors de l'atelier sur la production de métadonnées. Luxembourg. Document Eurostat 3331.

Australie, Bureau of Statistics (2005). Labour statistics: concepts, sources and methods. Canberra, Statistical Concepts Library. Disponible à l'adresse : *www.abs.gov.au/AUSSTATS/abs@nsf/DirClassManually Catalogue/59D849DC7BOIFCCECA257/10FOOI F6E5B.* Open Document Catalogue n° 6102.0.55.001.

Backlund, S. (1996). Future directions on IT issues. Mission report to National Statistical Centre. République démocratique populaire lao, Vientiane.

Banque mondiale (1991). *The SDA survey instrument: an instrument to capture social dimensions of adjustment.* Washington. Poverty and Social Policy Division, Technical Department, Africa Division.

Brogan, D. (2003). *Comparison of Data Analysis Software Suitable for Surveys in Developing Countries,* Organisation des Nations Unies, Division de statistique, New York.

Bureau international du Travail (1990). *Enquête sur la population économiquement active, l'emploi, le chômage et le sous-emploi : manuel de concepts et méthodes du BIT*. Genève, Bureau international du Travail.

Chromy, J. et S. Abeysasekara (2003). *Analytical Uses of Survey Data.* Organisation des Nations Unies, Division de statistique, New York.

Chronholm, P. et Edsfeldt (1996). Course and seminar on systems design. Mission report to Central Statistics (CSS), Pretoria.

Giles, M. (1996). *Turning Data into Information: A Manual for Social Analysis*. Canberra, Australian Bureau of Statistics.

Glewwe, Paul (2005). Aperçu de la conception d'un questionnaire pour les enquêtes sur les ménages dans les pays en développement. *Enquêtes sur les ménages dans les pays en développement et les pays en transition*. Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Graubard, B. et E. Korn (2002). *The Use of Sampling Weights in the Analysis of Survey Data.* Organisation des Nations Unies, Division de statistique, New York.

Jambwa, M. et L. Olsson (1987). Application of database technology in the African context. Document présenté à la quarante-sixième session de l'International Statistical Institute, Tokyo.

Jambwa, M., C. Parirenyatwa et B. Rosen (1989). *Data processing at the Central Statistical Office: Lessons from recent history.* Central Statistics Office, Harare.

Lagerlöf, Birgitta (1988). *Development of systems design for national household surveys.* SCB R&D report, n° 4. Stockholm, Statistics Sweden.

Lehtonen, R. et E. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York, Wiley & Sons.

Lundell, L. (1996). *Information systems strategy for CSS.* Report to Central Statistical Service (CSS), Pretoria.

_____ (2003). *Census data processing experiences.* Report to Central Bureau of Statistics (CBS), Windhoek.

Macro International, Inc. (1996). *Sampling Manual,* DHS-III Basic Document n° 6. Calverton, Maryland, Macro International, Inc.

Munoz, Juan (2003). Guide pour la gestion des données d'enquêtes sur les ménages. *Enquêtes sur les ménages dans les pays en développement et les pays en transition.* Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Namibie, Bureau central de statistique (1996). *The 1993/1994 National Household Income and Expenditure Survey (NHIES).* Administrative and technical report, Windhoek, National Planning Commission.

Olofsson, P. (1985). Proposals for Survey Design, Kingdom of Lesotho, Report on short-term mission on a labour-force survey to Bureau of Statistics, Maseru.

Olsson, Ulf (1990)a. *Approaches to agricultural statistics in developing countries: an appraisal of ICO's experiences,* n° 12. Stockholm, SCB (Statistics Sweden, International Consulting Office). 20 juillet.

_____ (1990)b. *Applied statistics lecture notes: special reports.* TAN 1990:1. Stockholm Statistics Sweden International Consulting Office.

Organisation des Nations Unies (1982). *Programme de mise en place de dispositifs nationaux d'enquêtes sur les ménages : traitement des données d'enquête : problèmes et procédures.* DP/UN/INT-81-041/1. New York, Organisation des Nations Unies, Département de la coopération technique pour le développement et Bureau de statistique.

_____ (1985). *Programme de mise place de dispositifs nationaux d'enquêtes sur les ménages : household income expenditure surveys: a technical study.* DP/UN/INT.88-X01/6E. New York, Organisation des Nations Unies, Département de la coopération technique pour le développement, Bureau de statistique.

Pettersson, Hans (2005). Conception de cadres directeurs d'échantillonnage et d'échantillons-maîtres pour les enquêtes sur les ménages dans les pays en développement. *Enquêtes sur les ménages dans les pays en développement et les pays en transition.* Études méthodologiques, n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Puide, Annika (1995). Report on a mission to Takwimu, Dar es Salaam, 21 novembre-21 décembre 1994. TANSTAT 1994, 20 (20 janvier 1995). Stockholm, Statistics Sweden, International Consulting Office.

Rauch, L. (2001). Best Practices in Designing Websites for Dissemination of Statistics. Conference of European Statisticians Methodological Material. Genève, Organisation des Nations Unies, Commission de la statistique et Commission économique pour l'Europe.

Rosen, B. et B. Sundgren (1991). *Documentation for re-use of microdata from surveys carried out by Statistics Sweden,* Working paper for Research and Development Unit, Statistics Sweden, Stockholm.

_____ (2002)a. Mission on sampling: framework for the master sample, Kingdom of Lesotho. Report from a mission to the Bureau of Statistics, Maseru, Lesotho, 1er-15 juin 2002. LESSSTAT 2002:7. Stockholm: Statistics Sweden, International Consulting Office.

_____ (2002)b. *Report on the short-term mission on estimation procedure for master sample surveys.* Maseru, Bureau of Statistics, Kingdom of Lesotho.

Rosen, Beugt (1991). *Estimation in the income, consumption and expenditure survey,* ZIMSTAT 1991: 8:1.

Shah, B., B. Barnwell et G. Bieler (1996). *SUDAAN User Manual: Release 7.0,* Research Triangle Park, Caroline du Nord. Research Triangle Institute.

Silva, P. Pedro Luis do Nascimento (2005). Établissement de rapports et compensation des erreurs autres que les erreurs d'échantillonnage au Brésil : pratique actuelle et défis pour l'avenir. *Enquêtes sur les ménages dans les pays en développement et les pays en transition.* Études méthodologiques n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

SPSS, Inc. (1988). *SPSS/PC+V2.0 Base Manual*. Chicago, Illinois, SPSS.

St. Catherine, Edwin (2003). Review of data processing, analysis and dissemination for *Designing Household Survey Samples: Practical Guidelines.* Réunion du Groupe d'experts des Nations Unies chargés de revoir

le projet de manuel sur la conception d'enquêtes par sondage sur les ménages. New York, 3-5 décembre 2003.

Sundgren, B. (1984). *Conceptual Design of Databases and Information Systems,* P/ADB Report E19. Stockholm, Statistics Sweden.

_____ (1986). *User-Oriented Systems Development at Statistics Sweden*. U/ADB Report E24, Stockholm, Statistics Sweden.

_____ (1991). *Une architecture des systèmes d'information destinée aux organismes nationaux et internationaux de statistique : lignes directrices et recommandations*. Genève, Organisation des Nations Unies, Commission de statistique et Commission économique pour l'Europe.

_____ (1995). *Guidelines: Modelling Data and Metadata*. Genève, Organisation des Nations Unies, Commission de statistique et Commission économique pour l'Europe.

Svensson, R. (1996). *The Census Data Entry Application.* Report from a mission to Central Statistical Service (CSS), Pretoria.

Thiel, Lisa Olson (2001). *Designing and developing a web site.* Report from a mission to Bureau of Statistics, Maseru, Lesotho, 12-23 novembre 2001. LESSTAT: 2001:17. 28 décembre. Stockholm, SCB Statistics Sweden, Interntional Consultancy Office.

Verma, Vijay (1982). The estimation and presentation of sampling errors, World Fertility Survey, *Technical Bulletins* n° 11 (décembre). La Haye, International Statistical Institute. Voorburg (Pays-Bas).

Wallgren, Anders *et al.* (1996). *Graphing Statistics and Data: Creating Better Charts.* Thousand Oaks, California, Sage Publications, Inc.

Yansaneh, I. (2005). Aperçu des problèmes de conception d'échantillons pour les enquêtes sur les ménages dans les pays en développement et les pays en transition. Organisation des Nations Unies, Division de statistique, New York. *Enquêtes sur les ménages dans les pays en développement et les pays en transition.* Études méthodologiques n° 96. Publication des Nations Unies, numéro de vente : F.05.XVII.6.

Annexe I

Éléments essentiels de la conception de l'échantillon

A.1. Introduction

1. Le sondage est une technique qui consiste à sélectionner une partie de la population et à généraliser les résultats obtenus au moyen de cette fraction à l'ensemble de la population dont l'échantillon a été sélectionné. D'une manière générale, il y a deux types d'échantillons, à savoir les échantillons probabilistes et non probabilistes. Dans tout ce guide, l'accent a été mis sur les échantillons probabilistes. L'aperçu ci-après, accompagné d'exemples, portera sur les unités d'échantillonnage, la conception des échantillons et les principales stratégies d'échantillonnage.

A.2. Unités et concepts

2. Il y a lieu de commencer en définissant les unités et les concepts les plus communément utilisés pour les enquêtes par sondage. *Élément :* les éléments ou unités d'une population sont les unités à propos desquelles l'on cherche à rassembler des informations. Il peut s'agir d'unités élémentaires qui constituent la population au sujet de laquelle des généralisations doivent être faites. Par exemple, dans le contexte d'une enquête sur la fécondité, les éléments ultimes sont habituellement les femmes en âge de procréer. Pour pouvoir plus facilement rassembler des données, il est absolument essentiel que les éléments soient définis et puissent être identifiés facilement.

3. *Population :* La population est la somme des éléments définis ci-dessus. Les éléments sont par conséquent les unités fondamentales qui constituent et définissent la population. Il est essentiel de définir la population en termes de :

- Contenu, ce qui signifie qu'il faut définir le type et les caractéristiques des éléments qui constituent la population;
- Étendue, c'est-à-dire les limites géographiques du territoire couvert par l'enquête;
- Temps, c'est-à-dire la période temporelle à laquelle se rattache la population.

4. *Unités d'observation :* Il s'agit des unités qui font l'objet des observations. Dans le cas des enquêtes menées par entrevues, ces unités sont appelés *déclarants.* Les déclarants sont les éléments qui fournissent les informations demandées. Il y a lieu de noter à ce propos que, dans certains cas, les unités d'observation et les déclarants peuvent être différents. Par exemple, si l'enquête porte sur les enfants de moins de 5 ans, ce seront normalement les parents qui fourniront les informations requises concernant leurs enfants. En pareil cas, les enfants sélectionnés dans l'échantillon sont les unités observation tandis que ce sont les parents qui sont les déclarants.

5. *Unités d'échantillonnage :* Les unités d'échantillonnage sont utilisées pour sélectionner les éléments à inclure dans l'échantillon. Dans un échantillonnage par élément, chaque unité d'échantillonnage ne contient qu'un seul élément, tandis que, dans le cas d'un échantillonnage en grappes,

par exemple, une unité d'échantillonnage se compose d'un groupe d'éléments appelé *grappe.* Par exemple, une zone d'énumération (ZE), en tant qu'unité d'échantillonnage élémentaire, contiendrait une grappe de ménages. Il se peut que des unités d'échantillonnage différentes soient utilisées pour la même enquête. Un bon exemple est celui de l'échantillonnage à phases multiples, qui utilise une hiérarchie d'unités d'échantillonnage (voir le chapitre 3).

6. *Unités sélectionnées :* Les unités d'échantillonnage retenues peuvent être appelées unités sélectionnées, et les valeurs des caractéristiques des unités sélectionnées à propos desquelles l'on cherche à obtenir des informations sont appelées unités observées. *Unité d'analyse :* Il s'agit de l'unité utilisée au stade de la tabulation et de l'analyse. Il pourra s'agir d'une unité élémentaire ou d'un groupe d'unités élémentaires. Comme indiqué ci-dessus, il y a lieu de noter que l'unité d'analyse et l'unité déclarante ne seront pas nécessairement les mêmes.

7. *Cadre d'échantillonnage :* Le cadre d'échantillonnage sert à identifier et à sélectionner les unités d'échantillonnage qui feront partie de l'échantillon ainsi qu'à établir des estimations sur la base de données recueillies par sondage. Cela signifie que la population parmi laquelle l'échantillon doit être sélectionné doit être représentée sous une forme physique. Idéalement, le cadre doit être constitué d'unités d'échantillonnage appartenant toutes à la population étudiée et étant toutes identifiées comme il convient. Les cadres doivent être exhaustifs et, de préférence, s'exclure mutuellement (pour de plus amples détails, voir le chapitre 4). Les types de cadres les plus communément utilisés pour les enquêtes sont les listes, le cadre géographique et les cadres multiples.

8. *Listes :* Une liste énumère les unités d'échantillonnage dans lesquelles un échantillon peut être sélectionné directement. Il est préférable que la liste contienne des informations pertinentes et exactes au sujet de chaque unité d'échantillonnage, comme sa taille et ses autres caractéristiques. Cette information supplémentaire facilite la conception et/ou la sélection d'échantillons efficaces.

9. *Cadres géographiques :* Les cadres géographiques sont des cadres à phases multiples qui sont, d'une manière générale, communément utilisés dans les enquêtes sur les ménages. Le cadre se compose d'unités géographiques sélectionnées en une ou plusieurs phases. Dans le cas d'une conception en deux phases, par exemple, le cadre se composera de grappes, qui peuvent être appelées unités primaires d'échantillonnage (UPE); pour les UPE sélectionnées, une liste des ménages deviendra le cadre secondaire. D'une façon générale, des cadres sont nécessaires pour chacune des phases de la sélection. La durabilité du cadre diminue à mesure que l'on descend dans la hiérarchie des unités.

10. *Unités géographiques :* Les unités géographiques couvrent les territoires spécifiés définis par des limites clairement déterminées, qui peuvent être des caractéristiques physiques comme routes, rues, cours d'eau, voies ferrées, ou des limites imaginaires représentant les limites officielles entre circonscriptions administratives. Les zones d'énumération du recensement sont habituellement choisies de manière à correspondre aux circonscriptions administratives les plus restreintes qui existent dans le pays, ce qui facilite l'addition des dénombrements opérés au niveau des circonscriptions administratives, en tant que domaines.

11. Le *cadre* ou les cadres utilisés pour une enquête sur les ménages doivent permettre d'avoir accès à toutes les unités d'échantillonnage de la population visée de sorte que chaque unité ait une probabilité de sélection connue autre que zéro. Pour accéder à la population, il est prélevé un échantillon à partir du cadre, habituellement en deux ou plusieurs phases de sélection. Le cadre utilisé pour la première phase de l'échantillonnage doit comprendre toutes les unités d'échantillonnage désignées. Lors des phases ultérieures, les cadres de sélection de l'échantillon servent uniquement à faire un

choix parmi les unités d'échantillonnage sélectionnées lors de l'étape précédente. Le cadre d'échantillonnage peut être stocké sur support papier et/ou par des moyens électroniques.

A.3. Conception de l'échantillon

12. D'une façon générale, l'on entend par conception la sélection et l'estimation de l'échantillon. Il s'agit par conséquent de savoir comment peut être sélectionnée la partie de la population qui fera l'objet de l'enquête. Dans la pratique, la conception consiste à déterminer la taille et la structure de l'échantillon, compte tenu des coûts de l'enquête. La conception privilégiée est celle qui donne la meilleure précision possible dans des limites déterminées de coût, ou qui permet de réaliser l'enquête aux moindres frais à un niveau de précision spécifié.

13. Il y a lieu de souligner d'emblée, toutefois, que la sélection de l'échantillon ne peut pas être dissociée des autres aspects de la conception et de la réalisation de l'enquête. Ainsi, la théorie de l'échantillonnage a pour but de déterminer comment, pour une population donnée, les estimations provenant de l'enquête et les erreurs d'échantillonnage connexes sont liées à la taille et à la structure de l'échantillon.

A.3.1. Conditions préalables à la conception d'un échantillon probabiliste

- La population cible doit être clairement définie.
- En cas d'échantillonnage à plusieurs phases, il peut être établi un ou plusieurs cadres d'échantillonnage.
- Les objectifs de l'enquête doivent être clairement définis en termes de contenu, de variables d'analyse et de niveaux de désagrégation (par exemple, des estimations ou des données doivent-elles être obtenues aux échelons national, rural/urbain, provincial, de district ?).
- Les contraintes budgétaires et les contraintes liées au travail de terrain doivent être prises en considération.
- Le degré de précision requis doit être clairement défini pour déterminer la taille de l'échantillon.

A.3.2. Avantages de l'échantillonnage probabiliste dans le cas d'enquêtes de grande envergure sur les ménages

- Il permet de sélectionner l'échantillon parmi l'ensemble de la population cible.
- Il réduit la distorsion due à l'effet d'échantillonnage.
- Il permet de généraliser et d'étendre les résultats obtenus au sujet de l'échantillon à l'ensemble de la population parmi laquelle l'échantillon a été sélectionné.
- Il permet de calculer les erreurs d'échantillonnage, qui sont des mesures de fiabilité.
- Il permet à l'enquêteur, selon certains, de présenter les résultats sans devoir justifier l'utilisation de méthodes non scientifiques.

A.3.3. Procédures de sélection, de réalisation et d'estimation

- Chaque élément de la population doit être représenté dans le cadre dont est tiré l'échantillon.

- La sélection de l'échantillon doit être fondée sur un processus aléatoire donnant à chaque unité une probabilité de sélection spécifiée.
- Toutes les unités sélectionnées, et seulement elles, doivent être recensées.
- Pour estimer les paramètres de la population constituée par l'échantillon, les données relatives à chaque unité/élément doivent être pondérées conformément à sa probabilité de sélection.

14. La sélection aléatoire des unités réduit le risque que l'échantillon ne soit pas représentatif. Ce type de sélection est donc une garantie qui permet d'éliminer les effets des causes imprévues de distorsion. La méthode utilisée pour sélectionner l'échantillon dépend du système d'échantillonnage employé. Plus la conception de l'échantillon est complexe, et plus exigeantes seront les procédures de sélection.

A.4. Éléments essentiels des stratégies d'échantillonnage probabiliste

15. Il existe plusieurs méthodes d'échantillonnage probabiliste pouvant être utilisées pour concevoir un échantillon, notamment les méthodes d'échantillonnage aléatoire simple, d'échantillonnage systématique, d'échantillonnage stratifié et d'échantillonnage en grappes. Chacune de ces méthodes sera discutée brièvement ci-dessous, avec quelques exemples.

A.4.1. Échantillonnage aléatoire simple

16. L'échantillonnage aléatoire simple (SRS) est une méthode de sélection probabiliste selon laquelle chaque élément de la population a une probabilité égale de sélection. La sélection de l'échantillon peut se faire avec ou sans remplacement. Cette méthode est rarement utilisée pour des enquêtes de grande envergure sur les ménages, car elle est coûteuse en raison des listes qui doivent être établies et des déplacements nécessaires. Elle peut être considérée comme la plus simple des formes d'échantillonnage probabiliste applicable aux situations dans lesquelles l'on ne dispose pas déjà d'informations concernant la structure de la population cible. L'échantillonnage aléatoire simple est une méthode attrayante car les procédures de sélection et d'estimation (par exemple des erreurs d'échantillonnage) sont peu complexes.

17. Bien que la méthode d'échantillonnage aléatoire simple ne soit guère utilisée, elle est importante pour la théorie de l'échantillonnage, essentiellement en raison de la simplicité de ses propriétés mathématiques. La plupart des théories et méthodes statistiques, par conséquent, prennent pour hypothèse une sélection aléatoire simple des éléments. En fait, toutes les autres sélections d'échantillons probabilistes peuvent être considérées comme des restrictions à la méthode d'échantillonnage aléatoire simple qui éliminent certaines combinaisons d'éléments de la population. L'échantillonnage aléatoire simple a un double but :

- Il constitue un point de référence qui permet de comparer l'efficacité relative des autres méthodes d'échantillonnage.
- Il peut être utilisé comme la méthode finale de sélection des unités élémentaires dans le contexte de conceptions plus complexes, comme l'échantillonnage en grappes à plusieurs phases et l'échantillonnage stratifié.

Les exemples ci-dessous illustrent le calcul de la probabilité de sélection dans le cas d'un échantillonnage aléatoire simple :

1. Nous prenons tout d'abord une population finie de 100 ménages $H1$, $H2$, ... H_i, ... $H100$ ayant des niveaux de revenu $X1$, $X2$, ... X_i, ... $X100$
   Dans cet exemple, la probabilité de sélection d'une unité déterminée est $\frac{1}{100}$.

2. Deuxième exemple : pour prélever un échantillon de ménages, les ménages cibles peuvent être énumérés l'un après l'autre dans un cadre ou une liste, et il peut être sélectionné au hasard un échantillon de 25 d'entre eux. Selon la méthode de sélection à probabilité égale (EPSEM), f est la fraction d'échantillonnage globale.
   Ainsi, $f = \frac{n}{N}$.
   Si $n = 25$ est la taille de l'échantillon et $N = 100$, le nombre total de ménages, la fraction d'échantillonnage, qui constitue la probabilité de sélection, est égale à :
   $\frac{25}{100} = \frac{1}{4}$.

A.4.1.1. Types de sélection d'un échantillon selon la méthode de l'échantillonnage aléatoire simple

18. Deux méthodes de sélection d'échantillons sont communément utilisées dans le cadre d'un échantillon aléatoire simple, à savoir :

a) L'échantillonnage aléatoire simple avec remplacement (SRSWR);

b) L'échantillonnage aléatoire simple sans remplacement (SRSWOR).

Échantillonnage aléatoire simple avec remplacement.

19. L'échantillonnage aléatoire simple avec remplacement est une méthode de sélection aléatoire d'un échantillon parmi une population, qui consiste à remplacer l'élément choisi de la population à l'issue de chaque étape. La probabilité de sélection d'un élément demeure inchangée après chaque étape, et les échantillons indépendants sélectionnés sont indépendants les uns des autres. Cette propriété explique pourquoi l'échantillonnage aléatoire simple est utilisé comme méthode d'échantillonnage par défaut dans nombre d'études statistiques théoriques. En outre, comme l'hypothèse à la base de la méthode d'échantillonnage aléatoire simple simplifie considérablement les formules à utiliser pour les estimations, comme les estimations de la variance, cette méthode est utilisée comme point de référence. L'on trouvera au paragraphe 20 ci-dessous les formules à appliquer pour estimer la moyenne (A.1) et la variance (A.2) de la moyenne de l'échantillon dans le cas d'un échantillonnage simple avec remplacement. Les formules sont illustrées par des exemples numériques.

20. Pour un échantillon de n unités sélectionnées au moyen de la méthode de l'échantillonnage simple avec remplacement, pour lesquelles il a été rassemblé des informations concernant la variable x, la moyenne et la variance sont données par les formules suivantes :

1. *Moyenne*

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i = \frac{1}{n}[x_1 + x_2 + \ldots \ldots + x_n] \qquad \textbf{(A.1)}$$

Si $x_1 = 24,\ x_2 = 30,\ x_3 = 27,\ x_4 = 36,\ x_5 = 31,\ x_6 = 38,\ x_7 = 23,\ x_8 = 40,\ x_9 = 25,\ x_{10} = 32,$
on a : $\bar{x} = \frac{24 + 30 + 27 + \ldots \ldots + 25 + 32}{10} = 30{,}6.$

2. *Variance*

$$V(\bar{x}) = \frac{s^2}{n} \qquad \textbf{(A.2)}$$

$$\text{où } s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_i^n x_i^2 - \frac{x^2}{n}\right] = \frac{1}{n-1}\left(\sum x_i{}^2 - n\bar{x}^{\,2}\right) \qquad \textbf{(A.3)}$$

$x^2 = (\sum x_i)^2 = 93\ 636.$

Par un calcul sur la base de ces valeurs,

$$s^2 = \frac{(9\ 684 - 9\ 364)}{9} = 35{,}56$$

$$V(\bar{x}) = \frac{35{,}56}{10} = 3{,}56$$

$\text{Se}(\bar{x}) = \sqrt{3{,}56}.$

Échantillonnage aléatoire simple sans remplacement

21. Il est préférable, à première vue, d'utiliser une méthode d'échantillonnage sans remplacement car on obtient ainsi plus d'informations, étant donné qu'il ne peut pas y avoir de répétition des unités d'échantillonnage. La stratégie d'échantillonnage aléatoire simple sans remplacement est donc la procédure d'échantillonnage aléatoire simple la plus fréquemment utilisée. Selon cette méthode, le processus de sélection est poursuivi jusqu'à ce qu'il soit sélectionné *n* unités distinctes, toutes les répétitions étant ignorées. Cela revient à conserver l'unité ou les unités sélectionnées ou à en sélectionner une autre, sur la base d'une probabilité égale, parmi les unités restantes de la population.

Certaines des propriétés de la méthode d'échantillonnage aléatoire simple sans remplacement sont les suivantes :

- Elle donne un échantillon de taille fixe;
- Elle aboutit à une probabilité de sélection égale pour chaque élément ou unité (EPSEM);
- Comme dans le cas de l'échantillonnage aléatoire simple avec remplacement, la moyenne et la variance de l'échantillon sont des estimations non biaisées des paramètres de la population.

22. L'on trouvera au paragraphe 24 ci-après les formules utilisées pour estimer la moyenne et la variance dans le cas d'un échantillonnage aléatoire simple sans remplacement (A.4 et A.5). L'on trouvera également des exemples numériques sur la méthode à suivre pour calculer la moyenne et la variance de l'échantillon.

23. Supposons que le nombre total d'écoles primaires d'une région soit de 275. Il en est sélectionné sans remplacement un échantillon de 55. Les chiffres ci-dessous sont le nombre d'employés (y_i) de chacune des écoles sélectionnées.

5	10	32	6	8	2
15	16	35	7	50	6
2	6	47	20	20	6
7	6	35	6	16	2

21	2	48	4	15	2
7	5	46	6	7	
4	4	8	2	6	
7	2	7	8	2	
5	12	10	6	2	
2	40	7	7	19	

$\sum y_i = 688$, nombre total d'employés

$\sum y_i^2 = 18\,182$

1. La moyenne de l'échantillon est :

$$\bar{y} = \frac{\sum y_i}{n} \qquad \text{(A.4)}$$

où n est la taille de l'échantillon.

Sur la base d'un calcul avec ces chiffres,

$\bar{y} = \frac{688}{55} = 12{,}5$

2. La variance de la moyenne de l'échantillon est :

$$V(\bar{y}) = 1 - f\frac{s_y^2}{n} \qquad \text{(A.5)}$$

où $1 - f$ est le facteur de correction de la population et

$$s_y^2 = \frac{1}{n-1}\left[\sum y_i^2 - n\bar{y}^2\right] = \frac{1}{54}[18\,182 - 8\,594] = 177{,}56 \qquad \text{(A.6)}$$

Donc $V(\bar{y}) = \left(1 - \frac{55}{275}\right)177{,}56/55 = 2{,}58$ et $\text{Se}(\bar{y}) = \sqrt{2{,}58}$

A.4.2. Échantillonnage systématique

24. L'échantillonnage systématique est une méthode de sélection d'un échantillon probabiliste selon laquelle l'échantillon est obtenu en sélectionnant chaque $k^{ième}$ élément de la population, k étant un chiffre supérieur à l'unité. Le premier élément de l'échantillon doit être sélectionné au hasard parmi les k premiers éléments. La sélection est faite sur une liste ordonnée. Il s'agit d'une méthode de sélection fréquemment utilisée, surtout lorsque les unités sont nombreuses et sont numérotées dans l'ordre de 1 à N. Supposons que N, le nombre total d'unités, soit un multiple intégral de la taille requise de l'échantillon n et que k soit un intégrant de sorte que $N = nk$. Il est alors sélectionné au hasard un chiffre compris entre 1 et k. Supposons que 2 soit le début de la recherche aléatoire, de sorte que l'échantillon a une taille n dont les unités seront numérotées consécutivement comme suit :

$2, 2 + k, 2 + 2k, \ldots \ldots, 2 + (n - 1)K$

Il y a lieu de noter que l'échantillon comprend la première unité sélectionnée au hasard et chaque *k* unité, jusqu'à ce que l'on obtienne l'échantillon de la taille requise. L'intervalle *k* divise la population en grappes ou groupes. Selon cette procédure, nous sélectionnons une grappe d'unités sur la base d'une probabilité 1/*k*. Comme le premier chiffre est choisi au hasard entre 1 et *k,* chaque unité des grappes censément égales a la même probabilité de sélection, 1/*k*.

A.4.2.1. Échantillonnage systématique linéaire

25. Si *N,* le nombre total d'unités, est un multiple de la taille requise de l'échantillon, autrement dit si $N = nk$, où *n* est la taille requise et *k* est un intervalle d'échantillonnage, les unités de chacun des échantillons systématiques possibles sont égales à *n*. En pareille situation, le système équivaut à classer les *N* unités en échantillons de *n* unités chacun et à sélectionner une grappe sur la base d'une probabilité 1/*k*. Lorsque $N = nk$, $\bar{y}$ est l'estimateur non biaisé de la moyenne de la population $\bar{Y}$. D'un autre côté, lorsque *N* n'est pas un multiple de *n,* le nombre d'unités sélectionnées au moyen de la méthode systématique avec un intervalle d'échantillon *k* égale à l'intégrant le plus proche de *N*/*n* peut ne pas nécessairement être égal à *n.* Ainsi, lorsque *N* n'est pas égal à *nk,* les tailles de l'échantillon varieront et la moyenne de l'échantillon sera un estimateur biaisé de la moyenne de la population. La figure A.1 ci-dessous illustre la sélection de l'échantillon selon la méthode d'échantillonnage systématique linéaire.

Figure A.1
Échantillonnage systématique linéaire (*sélection de l'échantillon*)

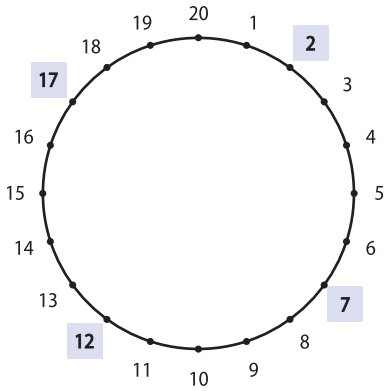
1	2	**3**	4	5	6	7	**8**	9	10
11	12	**13**	14	15	16	17	**18**	19	20

L'exemple ci-dessus illustre la sélection d'un échantillon de 4 parmi un groupe de 20 étudiants. Le chiffre pris comme point de départ de la sélection au hasard est 3, $N = 20$, $n = 4$ et $k = 5$. L'échantillon obtenu comprend les unités numérotées 3, 8, 13 et 18.

A.4.2.2. Échantillonnage systématique circulaire

26. Nous avons relevé que, selon la méthode d'échantillonnage systématique linéaire, la taille effective de l'échantillon variait par rapport à la taille souhaitée et que la moyenne de l'échantillon est un estimateur biaisé de la moyenne de la population lorsque *N* n'est pas un multiple de *n.* L'on peut cependant éliminer cette limitation en utilisant la méthode d'échantillonnage systématique circulaire. Selon cette méthode, les unités sont rangées en cercle, de sorte que la dernière soit suivie par la première. Il est choisi comme point de départ, au hasard, un chiffre compris entre 1 et *N* plutôt qu'entre 1 et *k.* La $k^{ième}$ unité est alors ajoutée jusqu'à ce que l'on ait sélectionné exactement *n* éléments. Lorsque l'on arrive à la fin de la liste, l'on reprend à partir du début. La figure A.2 illustre la sélection d'un échantillon selon la méthode d'échantillonnage systématique circulaire où $N = 20$, $n = 4$, $k = 5$ et le point de départ choisi au hasard est 7. Les unités sélectionnées sont par conséquent les unités 7, 12, 17 et 2.

Figure A.2
Sélection selon la méthode de l'échantillonnage systématique circulaire



A.4.2.3. Estimation dans le contexte de la méthode de l'échantillonnage systématique

27. Il est donné des formules pour estimer le total (A.7), la moyenne (A.8) et la variance (A.9) de l'échantillon ainsi que des exemples illustrant le calcul de la population estimative, de la moyenne de l'échantillon et de la variance.

1. Pour estimer le total, l'échantillon total est multiplié par l'intervalle de l'échantillonnage, de sorte que :

$$\hat{Y} = k\sum y_i \qquad \textbf{(A.7)}$$

L'estimation de la moyenne de la population est :

$$\bar{y} = k\frac{\sum y_i}{N} \qquad \textbf{(A.8)}$$

2. L'estimation de la variance est complexe en ce sens qu'il n'est pas possible de procéder à une estimation rigoureuse sur la base d'un seul échantillon systématique. Une solution consiste à prendre pour hypothèse que le numérotage des unités est aléatoire, auquel cas un échantillon systématique peut être considéré comme un échantillon aléatoire. L'estimation de la variance de la moyenne est par conséquent donnée par la formule :

$$V(\bar{y}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\sum s^2 \qquad \textbf{(A.9)}$$

où $s^2 = \frac{1}{n-1}\sum(y_i - y)^2$ et $y = \sum y_i$

28. Une estimation rigoureuse de la variance non biaisée d'un échantillon systématique peut être calculée en sélectionnant parmi une population déterminée plus d'un échantillon systématique.

Exemples numériques

29. Supposons qu'il y ait dans une province 180 exploitations commerciales ayant 30 têtes de bétail ou plus. Il est choisi un échantillon de 30 exploitations sur la base d'un échantillonnage systématique avec un intervalle de $k = 6$.

Le nombre de têtes de bétail (y_i) des 30 exploitations sélectionnées est indiqué ci-dessous.

60	200	45	50	40	79	35	41	30	120
300	65	111	120	200	42	51	67	32	40
46	55	250	100	63	90	47	82	31	50

et $\sum y_i = 2\ 542$

1. Le nombre estimatif de têtes de bétail est :

$\hat{Y} = k\sum y_i = 6 \times\ 2\ 542 = 15\ 252$

2. Le nombre estimatif moyen de têtes de bétail par exploitation est :

$$\bar{y} = k\frac{\left(\sum y_i\right)}{N} = 6 \times 2\ 542/180 = 84{,}7 \approx 85$$

3. La variance de la moyenne de l'échantillon, qui est calculée sur la base de l'hypothèse que la numérotation des exploitations est aléatoire, est :

$$V(\bar{y}) = 1 - f\frac{s_y^2}{n} \qquad \textbf{(A.10)}$$

où $s_y^2 = \frac{1}{n-1}\left\{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right\} = \frac{1}{29}(348\ 700 - 215\ 392{,}13) = 4\ 596{,}80$

de sorte que $V(\bar{y}) = (0{,}833)(153{,}227) = 127{,}64$

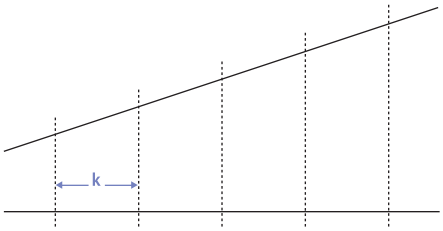
et Se $(\bar{y}) = \sqrt{127{,}64} = 11{,}3($

30. La méthode d'échantillonnage systématique présente plusieurs avantages et inconvénients.

a) Avantages :

- La sélection de la première unité dicte la composition de l'ensemble de l'échantillon, ce qui est de bon augure pour les opérations sur le terrain étant donné que des unités d'échantillonnage ultimes pourront être sélectionnées sur le terrain par les enquêteurs lorsqu'ils utiliseront une liste des unités;
- L'échantillon est réparti également sur l'ensemble de la population lorsque les unités du cadre sont numérotées comme il convient. Cependant, l'estimation de l'échantillon sera plus précise s'il se dégage une tendance de la population;
- L'échantillonnage systématique constitue une stratification implicite. La figure A.3 ci-après illustre la stratification implicite par une tendance linéaire monotonique.

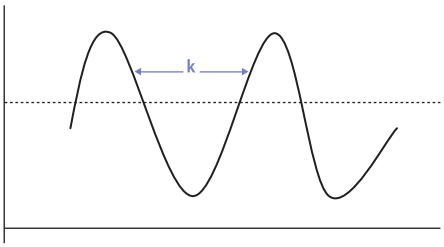
Figure A.3
Tendance linéaire monotonique



b) Inconvénients :

- S'il y a des variations périodiques dans la population, l'échantillonnage systématique peut donner des résultats qui sous-estiment ou surestiment la réalité. En pareil cas, l'intervalle d'échantillonnage coïncide avec les données. Par exemple, si l'on étudie la circulation pendant 24 heures dans une rue passante et que l'intervalle tombe pendant les heures de pointe, l'on aura toujours des chiffres élevés. L'étude donnera par conséquent des résultats qui seront des surestimations. La figure A.4 illustre une variation périodique qui peut contribuer à affecter la fiabilité des estimations dans le cas d'un échantillonnage systématique;
- À strictement parler, l'on ne peut pas obtenir d'estimation rigoureuse de la variance à partir d'un seul échantillon systématique;
- La méthode de sélection peut donner lieu à des abus de la part de certains enquêteurs ou agents de terrain.

Figure A.4
Fluctuations périodiques



A.4.3. Échantillonnage stratifié

31. Selon la méthode d'échantillonnage stratifié, les unités d'échantillonnage de la population sont divisées en groupes appelés strates. Il est habituellement procédé à une stratification de manière à subdiviser la population en groupes hétérogènes qui sont homogènes au plan interne. D'une manière générale, lorsque les unités d'échantillonnage sont homogènes du point de vue de la variable auxiliaire, appelée variable de stratification, la variabilité des estimateurs des strates se trouve habituellement réduite. Il convient de noter en outre que la stratification offre une flexibilité considérable en ce sens que les procédures d'échantillonnage et d'estimation peuvent varier d'une strate à l'autre.

32. Dans un échantillonnage stratifié, par conséquent, nous regroupons des unités/éléments qui sont plus ou moins semblables, de sorte que la variance δ_h^2 au niveau de chaque strate soit réduite. Dans le même temps, il est essentiel que les moyennes $(\bar{x}_h)$ des différentes strates soient aussi différentes que possible. L'on obtient une estimation appropriée pour l'ensemble de la population en combinant comme il convient les estimateurs par strates de la caractéristique à l'étude.

A.4.3.1. Avantages de l'échantillonnage stratifié

33. Le principal avantage de l'échantillonnage stratifié tient au fait qu'il permet d'améliorer la précision des estimations ainsi que d'utiliser des procédures d'échantillonnage différentes pour différentes strates. En outre, la stratification s'est avérée utile :

- Dans le cas de populations inégales, l'on peut utiliser des fractions d'échantillonnage importantes pour procéder à la sélection parmi un petit nombre d'unités plus nombreuses, ce qui donne plus de poids aux unités très nombreuses et, en définitive, réduit la variabilité de l'échantillonnage au niveau de chaque strate;
- Lorsque l'organisation chargée de l'enquête a plusieurs bureaux de terrain dans différentes régions correspondant aux divisions administratives du pays, auquel cas il peut être utile de considérer les régions comme des strates afin de réaliser plus facilement le travail sur le terrain;
- Quand les estimations doivent répondre à des marges d'erreurs spécifiées, non seulement pour l'ensemble de la population mais aussi pour certains sous-groupes comme les provinces, les populations urbaines ou rurales, le sexe, etc. De telles estimations peuvent aisément être obtenues au moyen d'une stratification;
- Si le cadre d'échantillonnage se présente sous forme de sous-cadres, qui peuvent correspondre à des régions ou à des catégories spécifiées d'unités, auquel cas il peut être commode du point de vue opérationnel et économique de considérer les sous-cadres comme des strates aux fins de la sélection de l'échantillon.

A.4.3.2. Récapitulation des étapes de la procédure d'échantillonnage stratifié

- L'ensemble de la population d'unités d'échantillonnage est subdivisé en sous-population homogène au plan interne mais hétérogène au plan externe.
- Il est sélectionné à l'intérieur de chaque strate un échantillon distinct prélevé parmi toutes les unités d'échantillonnage de la strate.
- À partir de l'échantillon obtenu dans chaque strate, il est calculé pour la strate considérée une moyenne (ou toute autre statistique) séparée. Les moyennes des strates, par exemple, sont alors pondérées comme il convient pour constituer une estimation combinée pour la moyenne de la population.
- Habituellement, l'on a recours à un échantillonnage proportionnel à l'intérieur de chaque strate lorsque les estimations globales — par exemple les estimations nationales — constituent le but de l'enquête et que celle-ci est transversale.
- L'on a recours à un échantillonnage disproportionné lorsque les domaines des sous-groupes sont prioritaires, par exemple lorsque l'on cherche à obtenir des estimations de fiabilité égale pour des régions infranationales.

A.4.3.3. Notations

34. Beaucoup de symboles et d'indices sont utilisés avec la méthode d'échantillonnage stratifié. Il faut par conséquent commencer par définir des notations et certains des symboles les plus communément utilisés dans cette stratégie d'échantillonnage.

Valeurs de la population

Pour H strates, le nombre total d'éléments de chaque strate sera dénoté par :
$N_1, N_2, \ldots \ldots N_h, \ldots \ldots N_H$.
Cette information est habituellement inconnue. La valeur de la population totale est :

$$\sum_{h}^{H} N_h = N \qquad \textbf{(A.11)}$$

Moyenne de la strate

$$\overline{X}_{hi} = \frac{1}{N}\sum_{i}^{N_h} X_{hi} = \frac{X_h}{N} \qquad \textbf{(A.12)}$$

où X_{hi} est la valeur du $i^{ième}$ élément de la $h^{ième}$ strate, et X_h est la somme de la $h^{ième}$ strate.

A.4.3.4. Pondérations

35. Les pondérations représentent généralement les proportions des éléments de la population que comportent les strates et :

$$W_h = \frac{N_h}{N} \qquad \textbf{(A.13)}$$

de sorte que $\sum W_h = 1$

$$S_h^2 = \frac{1}{N-1}\sum_{i=1}^{N_h}\left(X_{hi} - \overline{X}\right)^2 \qquad \textbf{(A.14)}$$

A.4.3.5. Valeurs de l'échantillon

36. Une valeur de l'échantillon est une estimation calculée à partir des *éléments sélectionnés* d'une strate. Dans la présente section, nous décrivons les symboles communément utilisés dans le contexte d'un échantillonnage stratifié :

a) Pour H strates, disons que les tailles de l'échantillon de chaque strate sont dénotées par n_1, $n_2, \ldots \ldots n_n$ où $\sum n_h = n$ est la taille de l'échantillon total;
b) Disons que x_{hi} est l'élément de l'échantillon i de la strate h;
c) Ainsi :

$\overline{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ est la moyenne de l'échantillon pour la strate h; **(A.15)**

d) Puis :

$\bar{x}_{st} = \sum W_h \bar{x}_h$ est la moyenne de l'échantillon global; **(A.16)**

e) Tandis que

$f_h = \frac{n_h}{N_h}$ est la fraction d'échantillonnage pour la strate *h*. **(A.17)**

La variance du $n_h^{ième}$ élément de la $h^{ième}$ strate est donnée par la formule :

$$v(\bar{x}_h) = \sum \left[1 - \frac{n_h}{N_h}\right] \frac{s_h^2}{n_h} \quad \textbf{(A.18)}$$

où s_h^2 est la variance de l'élément pour la $h^{ième}$ strate et est donnée par la formule :

$$s_h^2 = \frac{\sum (x_{hi} - \bar{x}_h)^2}{(n_h - 1)} \quad \textbf{(A.19)}$$

La variance de la moyenne de l'échantillon est donnée par la formule :

$$v(\bar{x}_{st}) = \sum W_h^2 (1 - f_h) \frac{s_h^2}{n_h} \quad \textbf{(A.20)}$$

Deux types de stratégies d'échantillonnage stratifié, à savoir la stratification proportionnelle et la stratification disproportionnée, sont exposés ci-après.

A.4.3.6. Stratification proportionnelle

37. Dans l'échantillonnage stratifié, l'allocation proportionnelle consiste à utiliser une fraction d'échantillonnage uniforme pour toutes les strates, ce qui signifie qu'il est sélectionné dans chaque strate la même proportion d'unités. Par exemple, si nous décidons de sélectionner un échantillon total de 10 %, nous devrons sélectionner 10 % des unités de chaque strate. Comme les taux d'échantillonnage sont les mêmes pour toutes les strates, les éléments sélectionnés varieront d'une strate à l'autre. À l'intérieur de chaque strate, la taille de l'échantillon sera proportionnelle au nombre d'éléments qu'elle comporte.

Dans ce cas, la fraction d'échantillonnage est donnée par la formule $f_h = \frac{n_h}{N_h} = \frac{n}{N}$, ce qui implique une conception EPSEM.

La moyenne de l'échantillon est $\bar{x}_{st} = \sum W_h \bar{x}_h$ **(A.21)**

La variance de la moyenne globale est $v(\bar{x}_{st}) = \frac{(1-f)}{n} \sum W_h s_h^2$ **(A.22)**

A.4.3.7. Stratification disproportionnée

38. La méthode d'échantillonnage disproportionné consiste à utiliser des taux d'échantillonnage différents pour les différentes strates, le but étant d'allouer à chacune d'elles des taux d'échantillonnage tels qu'on peut obtenir la moindre variance au coût unitaire moyen global.

39. Selon cette méthode, le taux d'échantillonnage dans une strate déterminée est proportionnel à l'écart type relatif à la strate considérée. Cela signifie que le nombre d'unités d'échantillonnage à sélectionner dans chaque strate dépendra non seulement du nombre total d'éléments, mais aussi de l'écart type de la variable auxiliaire.

Selon l'allocation disproportionnée, la notion de coût est également prise en considération. Par exemple :

$$C = C_o + \sum c_h n_h \qquad \text{(A.23)}$$

où C_o est le coût fixe et c_h le coût de la couverture de l'échantillon dans une strate donnée.
Fréquemment, nous pouvons tenir pour acquis que c_h est une constante pour toutes les strates. L'une des formules plus communément utilisées pour l'allocation disproportionnée des échantillons en strates est l'allocation de Neyman.
Lorsque c_h est une constante et que $\sum n_h$, la taille de l'échantillon global est fixe.
Le nombre d'unités devant être sélectionnées à l'intérieur de chaque strate est donné par la formule :

$$n_h = \frac{W_h s_h n}{\sum W_h s_h} \quad \text{or} \quad n_h = \frac{N_h s_h \cdot n}{\sum N_h s_h} \qquad \text{(A.24)}$$

La variance est donnée par :

$$v\left(\bar{x}_{st}\right) = \frac{\left(\sum W_h s_h\right)^2}{n} - \frac{1}{N}\sum W_h s_h^2 \qquad \text{(A.25)}$$

Le terme se trouvant à droite du signe moins est un facteur de correction de la population finie qui peut être supprimé si l'échantillon est sélectionné parmi une population très nombreuse, c'est-à-dire si la fraction d'échantillonnage est réduite.

A.4.3.8. Observations générales

- Les valeurs de la population S_h et C_h sont généralement inconnues, de sorte que des estimations peuvent être tirées d'enquêtes précédentes ou d'enquêtes pilotes.
- L'allocation disproportionnée n'est pas une méthode très efficace pour la sélection de proportions.
- Il peut y avoir des conflits entre les variables à optimiser dans le cas d'enquêtes transversales.
- D'une manière générale, l'allocation disproportionnée débouche sur la variance la plus réduite.

40. Les exemples ci-après illustrent le calcul de tailles d'échantillons et de variances selon les méthodes de stratification proportionnelle et disproportionnée. Dans cet exemple hypothétique, les écoles sont stratifiées sur la base du nombre d'employés. Le nombre total d'écoles primaires d'une province est de 275, et il est sélectionné et stratifié un échantillon de 55 écoles sur la base du nombre d'employés.

Tableau A.1
Nombre d'écoles, par effectif des employés

Strate	Effectifs des employés par école sélectionnée (y_{hi})	Nombre total d'écoles dans chaque strate (N_h)	Nombre d'écoles sélectionnées par strate		W_h	s_h^2	S_h	$W_h s_h$	$W_h s_h^2$
			Allocation proportionnelle (n_h)	Allocation disproportionnée (n_h)					
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	2,4,2,2,4, 2,2,4,2,2, 2,2,2,2,5,5	80	16	8	0,2909	1,663	1,289	0,3750	0,48
2	7,7,7,6,8, 7,7,6,7,6 6,8,6,7,8, 6,7,6,6,6	100	20	6	0,3636	0,537	0,733	0,2665	0,19
3	10,12,10,15, 21,16,20,20, 16,19,15	55	11	18	0,2000	15,564	3,945	0,7890	3,11
4	32,35,35,48, 46,47,50,40	40	8	23	0,1455	48,836	6,989	1,0169	7,10
		275	55	55	1,0000			2,4474	10,90

Note : N = nombre total d'écoles primaires
n = nombre total d'écoles primaires que comporte l'ensemble de l'échantillon
N_h = taille de la strate
n_h = taille de l'échantillon de la $h^{ième}$ strate.

A.4.3.9. Détermination de la taille de l'échantillon dans chaque strate

41. Se reporter au tableau A.1 ci-dessus.

Allocation proportionnelle

Dans le cas de l'allocation proportionnelle, c'est la fraction d'échantillonnage usuelle qui est employée.

Ainsi, $\frac{n}{N} = f$ est la fraction d'échantillonnage globale appliquée au nombre total d'unités que comporte la strate.

Dans l'exemple ci-dessus, $f = \frac{55}{275} = 0{,}2$ ou 20 %.

La répartition des tailles des échantillons est donnée dans la colonne 4 du tableau A.1; par exemple, la taille de l'échantillon pour la strate 1 est n_h = 0,2 × 80 = 16.

Allocation disproportionnée

La formule à employer pour obtenir les tailles des échantillons correspondant aux différentes strates est donnée par la formule :

$$n_h = \frac{W_h s_h}{\sum W_h s_h}(n) \qquad \textbf{(A.26)}$$

Par exemple, $n_h = \frac{0{,}3750}{2{,}4474} \times 55 = 8$ pour la strate 1.

Le reste des résultats est indiqué dans la colonne 5 du tableau.

A.4.3.10. Calcul des variances

42. Le calcul des variances selon les méthodes de stratification proportionnelle et disproportionnée est illustré par l'application des formules A.27 et A.28, respectivement.

Stratification proportionnelle

$$V\left(\bar{y}_{prop}\right) = \frac{1-f}{n}\sum w_h s_h^2 = \frac{(1-0{,}2)}{55}(10{,}9) = 0{,}16 \qquad \textbf{(A.27)}$$

Stratification disproportionnée

$$V\left(\bar{y}_{opt}\right) = \frac{\left(\sum w_h s_h\right)^2}{n} - \frac{1}{N}\sum w_h s_h^2 = \frac{(2{,}4474)^2}{55} - \frac{10{,}9}{275} = 0{,}07 \qquad \textbf{(A.28)}$$

A.4.3.11. En général

$$v\left(\bar{x}_{st}\right)OP \le v\left(\bar{x}_{st}\right)PROP \le v\left(\bar{x}_{st}\right) \le \left(\bar{x}_{st}\right)SRS \qquad \textbf{(A.29)}$$

A.4.4. Échantillonnage en grappes

43. Dans les sections précédentes, la discussion a porté sur les méthodes d'échantillonnage selon lesquelles les unités élémentaires d'échantillonnage étaient organisées de manière à constituer la liste à utiliser pour le cadre d'échantillonnage, de sorte que les différentes unités puissent être sélectionnées directement à partir de celui-ci. Dans le cas de l'échantillonnage en grappes, les unités de sélection de niveau plus élevé, par exemple les zones d'énumération (voir le chapitre 3), contiennent plus d'une unité élémentaire. Dans ce cas, l'unité d'échantillonnage est la grappe. Par exemple, une méthode simple de sélection d'un échantillon aléatoire de ménages d'une ville pourra consister à établir une liste de tous les ménages. Cela risque cependant de ne pas être possible étant donné que, dans la pratique, il n'y a sans doute pas de cadre complet englobant tous les ménages de la ville. Pour tourner ce problème, l'on peut constituer des grappes sous forme de pâtés de maisons et sélectionner ensuite un échantillon de pâtés de maisons et établir enfin une liste des ménages des pâtés de maisons sélectionnés. Si besoin est, l'on pourrait, dans chaque pâté de maisons, sélectionner un échantillon représentant par exemple 10 % des ménages.

A.4.4.1. Utilité de l'échantillonnage en grappes

44. Certaines des raisons qui militent en faveur d'une méthode d'échantillonnage en grappes, surtout dans le cas de conceptions à plusieurs phases, sont les suivantes :

- La mise en grappes réduit les frais de voyage et les autres coûts liés à la collecte des données;
- La mise en grappes permet d'améliorer la supervision, le contrôle, la couverture de suivi et d'autres aspects qui influent sur la qualité des données rassemblées;
- La construction du cadre est moins coûteuse dans la mesure où elle est réalisée par étapes. Par exemple, dans le cas d'un échantillonnage à phases multiples, comme discuté au chapitre 3, un cadre englobant l'ensemble de la population n'est requis que pour la sélection des UPE, c'est-à-dire des grappes lors de la première phase. Lors des phases suivantes, un cadre n'est requis que pour les unités sélectionnées lors de la phase précédente;
- En outre, les cadres d'unités plus nombreuses et de niveaux supérieurs tendent à être plus durables et peuvent par conséquent être utilisés sur des périodes plus longues. Des listes de petites unités comme des ménages et surtout des individus tendent à devenir obsolètes très rapidement;
- La mise en grappes facilite la réalisation de l'enquête sur le plan administratif.

45. D'une manière générale, il y a lieu de noter que, lorsque l'on compare un échantillon en grappes et un échantillon d'éléments de même taille, l'on constate que, dans le premier cas, le coût par élément est inférieur en raison du moindre coût de l'établissement d'une liste des éléments et/ou de leur localisation. D'un autre côté, la variance des éléments est plus forte en raison de l'homogénéité irrégulière des éléments (corrélation intraclasse) des grappes. L'on peut donner un exemple de la méthode fondamentale d'échantillonnage en grappes en prenant le cas d'une conception à une seule phase (les conceptions à phases multiples ont été présentées et discutées en détail au chapitre 3).

A.4.4.2. Échantillonnage en grappes à une seule phase

46. Il se peut que, dans un district déterminé, il ne soit pas possible d'obtenir une liste de tous les ménages pour en sélectionner ensuite un échantillon. Cependant, il se peut que l'on puisse trouver une liste de villages établie lors d'une enquête précédente ou tenue à des fins administratives. Nous obtiendrions alors un échantillon de villages, et ensuite des informations sur tous les ménages des villages sélectionnés. Cela serait une conception d'échantillonnage en grappes en une seule phase étant donné qu'après qu'un échantillon de villages a été sélectionné, toutes les unités de la grappe, en l'occurrence les ménages, sont prises en compte.

47. La sélection de l'échantillon, en cas de mise en grappes, peut être illustrée comme suit. Supposons qu'il soit sélectionné sur la base d'une probabilité égale un échantillon de la population des villages (grappes). Dans le cas d'un échantillonnage en grappes à une seule phase, tous les ménages des villages sélectionnés seraient inclus dans l'échantillon.

Comme

A = nombre total de villages

B = nombre total de ménages de la grappe

a = échantillon des villages

et comme par conséquent :

$aB = n$ représente le nombre d'unités élémentaires (ménages) de l'échantillon total

et

$AB = N$ est le nombre total de ménages de tous les villages,

la probabilité de sélection d'un élément sur la base d'une probabilité égale est donnée par la formule :

$$\frac{a}{A} \times \frac{B}{B} = \frac{n}{N} = f \qquad \text{(A.30)}$$

où N est le nombre total d'unités élémentaires et f la fraction d'échantillonnage. Dans ce cas, la probabilité de sélection est simplement $\frac{a}{A}$.

A.4.4.3. Formules à appliquer au calcul de la moyenne et de la variance de l'échantillon

48. L'on trouvera ci-dessous les formules à appliquer pour le calcul de la moyenne et de la variance de l'échantillon :

Moyenne de l'échantillon

$$\bar{y} = \frac{1}{aB} = \sum_{\alpha=1}^{\alpha} \sum_{\beta=1}^{\beta} \bar{y}_{\alpha\beta} = \frac{1}{a} \sum_{\alpha=1}^{\alpha} \bar{y}_{\alpha} \qquad \text{(A.31)}$$

La moyenne de l'échantillon est une estimation non biaisée de la moyenne de la population :

$$E(\bar{y}) = \frac{1}{A} \sum_{\alpha=1}^{a} \bar{y}_{\alpha} = \bar{Y} \qquad \text{(A.32)}$$

En fait, comme la taille de l'échantillon est fixe ($aB = n$) et que la sélection est basée sur une probabilité égale, la moyenne ($\bar{y}$) est une estimation non biaisée de la moyenne de la population $\bar{Y}$.

Variance

Si les grappes sont sélectionnées sur la base d'une sélection aléatoire simple, la variance peut être estimée comme suit :

$$V(\bar{y}) = (1 - f) s_{\alpha}^2 \qquad \text{(A.33)}$$

où $s_{\alpha}^2 = \frac{1}{a-1} \sum_{\alpha=1}^{a} (\bar{y}_{\alpha} - \bar{y})^2$

49. Il importe de noter que les valeurs sont exemptes d'erreur d'échantillonnage étant donné qu'elles sont fondées sur les valeurs de tous les éléments de B et non sur un échantillon. La variance de la moyenne de l'échantillon n'est due qu'aux variances entre les moyennes des grappes.

Annexe II

Liste des participants à la réunion du Groupe d'experts des Nations Unies chargés d'examiner le projet de manuel sur la conception des enquêtes sur les ménages, New York, 3-5 décembre 2003[1]

Nom	Titre et organisation
Oladejo Oyeleke Ajayi	Consultant statisticien (Nigéria)
Beverley Carlson	Division de la production, de la productivité et de la gestion, Commission économique pour l'Amérique latine et les Caraïbes, Santiago (Chili)
Samir Farid	Consultant statisticien (Égypte)
Maphion M. Jambwa	Conseiller technique, Communauté de développement de l'Afrique australe/Union européenne, Gaborone (Botswana)
Udaya Shankar Mishra	Associate Fellow, Université de Harvard, Boston, Massachusetts (États-Unis d'Amérique)
Jan Kordos	Professeur à la Faculté des sciences économiques de Varsovie (Pologne)
Edwin St. Catherine	Directeur du Bureau national de statistique (Sainte-Lucie)
Anthony Turner	Consultant spécialisé dans l'échantillonnage (États-Unis d'Amérique)
Shyam Upadhyaya	Directeur, Integrated Statistical Services (INSTAT) [Népal]
Ibrahim Yansaneh	Chef adjoint de la Division du coût de la vie, Commission de la fonction publique internationale, Organisation des Nations Unies, New York (États-Unis d'Amérique)

[1] Pour le rapport de la réunion du Groupe d'experts, voir le document ESA/STAT/AC.93/L.4.