

ENQUÊTE CONJOINTE CANADA/ÉTATS-UNIS SUR LA SANTÉ
GUIDE DE L'UTILISATEUR DU FICHIER DE MICRODONNÉES À
GRANDE DIFFUSION

Statistique Canada
United States National Center for Health Statistics

Juin 2004

Table des matières

| | |
|---|-----------|
| 1. INTRODUCTION..... | 1 |
| 2. CONTEXTE..... | 2 |
| 3. OBJECTIFS..... | 4 |
| 4. CONTENU DE L'ENQUÊTE..... | 5 |
| 5. PLAN D'ÉCHANTILLONNAGE..... | 7 |
| 5.1 POPULATION CIBLE..... | 7 |
| 5.2 PLAN D'ÉCHANTILLONNAGE ET RÉPARTITION DE L'ÉCHANTILLON..... | 7 |
| 5.3 ÉCHANTILLONNAGE DES MÉNAGES..... | 7 |
| 5.3.1 <i>Base de sondage au Canada</i> | 8 |
| 5.3.2 <i>Base de sondage aux États-Unis</i> | 8 |
| 5.4 ÉCHANTILLONNAGE DES RÉPONDANTS..... | 10 |
| 6. COLLECTE DES DONNÉES..... | 12 |
| 6.1 CONCEPTION DU QUESTIONNAIRE ET MÉTHODE DE COLLECTE DES DONNÉES..... | 12 |
| 6.1.1 <i>Traduction</i> | 13 |
| 6.2 INTERVIEW..... | 13 |
| 6.3 RÉDUCTION DE LA NON-RÉPONSE AU MINIMUM..... | 14 |
| 7. TRAITEMENT DES DONNÉES..... | 16 |
| 7.1 VÉRIFICATION..... | 16 |
| 7.2 CODAGE..... | 16 |
| 7.3 CRÉATION DE VARIABLES DÉRIVÉES ET GROUPEES..... | 17 |
| 7.4 PONDÉRATION..... | 17 |
| 7.5 SUPPRESSION DE RENSEIGNEMENTS CONFIDENTIELS..... | 17 |
| 8. PONDÉRATION..... | 19 |
| 8.1 AJUSTEMENTS APPLIQUÉS AU POIDS INITIAL..... | 19 |
| 9 QUALITÉ DES DONNÉES..... | 24 |
| 9.1 TAUX DE RÉPONSE..... | 24 |
| 9.1.1 <i>Taux de réponse au Canada</i> | 24 |
| 9.1.2 <i>Taux de réponse aux États-Unis</i> | 25 |
| 9.2 ERREURS DANS LES ENQUÊTES..... | 26 |
| 9.2.1 <i>Erreurs non dues à l'échantillonnage</i> | 26 |
| 9.2.2 <i>Erreurs d'échantillonnage</i> | 27 |
| 9.2.3 <i>Estimation de la variance</i> | 27 |
| 9.2.3.1 <i>Méthode du bootstrap pour l'estimation de la variance</i> | 28 |
| 9.2.3.1.1 <i>Programme Bootvar pour l'estimation de la variance</i> | 28 |
| 9.2.3.1.2 <i>Méthode du développement en série de Taylor (avec SUDAAN) pour l'estimation de la variance</i> | 28 |
| 9.2.3.2 <i>Méthode du développement en série de Taylor (avec SUDAAN) pour l'estimation de la variance</i> | 28 |
| 10. LIGNES DIRECTRICES POUR LA TOTALISATION, L'ANALYSE ET LA DIFFUSION..... | 31 |
| 10.1 LIGNES DIRECTRICES POUR L'ARRONDISSEMENT..... | 31 |
| 10.2 LIGNES DIRECTRICES POUR LA PONDÉRATION DE L'ÉCHANTILLON EN VUE DE LA TOTALISATION..... | 32 |
| 10.2.1 <i>Définitions : estimations de type nominal, estimations quantitatives</i> | 32 |
| 10.2.2 <i>Totalisation d'estimations de type nominal</i> | 33 |
| 10.2.3 <i>Totalisation d'estimations quantitatives</i> | 33 |
| 10.3 LIGNES DIRECTRICES POUR L'ANALYSE STATISTIQUE..... | 34 |

| | | |
|------------|--|-----------|
| 10.4 | LIGNES DIRECTRICES POUR LA DIFFUSION | 34 |
| 11. | UTILISATION DU FICHIER | 37 |
| 11.1 | UTILISATION DES POIDS | 37 |
| 11.2 | CONVENTION APPLIQUÉE POUR NOMMER LES VARIABLES | 37 |
| 11.2.1 | <i>Structure élémentaire des noms des variables de l'ECCES</i> | 37 |
| 11.2.2 | <i>Positions 1 et 2 : Nom des modules du questionnaire</i> | 38 |
| 11.2.3 | <i>Position 3 : Identificateur de l'ECCES</i> | 38 |
| 11.2.4 | <i>Position 4 : Cycle</i> | 39 |
| 11.2.5 | <i>Position 5 : Type de variable</i> | 39 |
| 11.2.6 | <i>Positions 6 à 8 : Nom de la variable</i> | 39 |
| 11.3 | ACCÈS AUX DONNÉES | 40 |

1. Introduction

L'Enquête conjointe Canada-États-Unis sur la santé (ECCES) est un projet collectif entrepris par la Division de la statistique de la santé de Statistique Canada et le « National Center for Health Statistics (NCHS) des Centers for Disease Control and Prevention » des États-Unis. Il consiste à réaliser une enquête téléphonique ponctuelle dans les deux pays, en contrôlant pour les effets du plan de sondage et de la mise en œuvre, en vue de produire des résultats comparables.

Le questionnaire élaboré conjointement par les deux organismes contient des questions sur les problèmes de santé chroniques, l'état fonctionnel, les déterminants de la santé et l'utilisation des services de santé. L'enquête a été réalisée auprès d'environ 3 500 ménages au Canada et 5 000 ménages aux États-Unis. Des plans d'échantillonnage par composition aléatoire (CA) comparables ont été utilisés au Canada et aux États-Unis et la collecte des données a été réalisée par les intervieweurs de Statistique Canada auprès des répondants américains et canadiens.

Le fichier de microdonnées à grande diffusion contient les données recueillies auprès des deux échantillons du 4 novembre 2002 au 31 mars 2003 (la collecte des données d'avril et de juin 2003 aux États-Unis a été prolongée de plusieurs semaines en avril et en juin 2003 – voir la section 6.2 pour plus de précisions) de façon représentative dans les dix provinces du Canada, les 50 États des États-Unis et le district de Columbia. Les données ont été recueillies auprès d'un adulte de 18 ans ou plus par ménage sélectionné parmi les personnes vivant dans des logements privés occupés. Ont été exclues de la base de sondage les personnes résidant en établissements, les membres à temps plein des Forces canadiennes ou américaines et les résidents des Territoires du Canada et des États-Unis.

Le présent document a pour but de faciliter l'analyse du fichier de microdonnées à grande diffusion de l'ECCES qui est décrit en détail dans le texte et les autres documents connexes.

Pour toute question concernant les ensembles de données ou leur utilisation, adressez-vous à :

Soutien technique ou renseignements généraux sur les données aux États-Unis :

Ligne d'information du NCHS : 1 (301) 458-INFO (4636) ou

Sans frais : 1 (866) 441-NCHS (6247)

Soutien technique ou renseignements généraux sur les données au Canada :

Section de l'accès aux données, Division de la statistique de la santé : 1 (613) 951-1653

Courriel : jcush-ecces@statcan.ca

2. Contexte

L'Enquête conjointe Canada/États-Unis sur la santé (ECES) a été créée à la suite d'une séance de discussion sur la façon de rendre les données plus comparables et intégrées qui a eu lieu durant l'échange États-Unis/Canada de 2000. Les échanges, qui consistent en une réunion de deux jours tenue alternativement dans la région de Washington DC et à Ottawa, ont débuté en 1999. L'objectif des réunions est de promouvoir la communication, la collaboration, la coopération, la synergie, la réalisation d'analyses comparatives et l'interaction entre les deux pays. Les participants à l'échange 2000 ont laissé entendre que le meilleur moyen de faire des comparaisons consisterait à réaliser une enquête conjointe dont le questionnaire, le plan de sondage, les méthodes de collecte des données, le traitement des données et la vérification seraient les mêmes et auraient lieu au même moment dans les deux pays. Une équipe formée d'employés des deux organismes a été créée pour explorer la faisabilité du projet et pour le mettre en œuvre.

L'ECES vise, grâce à l'intégration des plans de sondage et des méthodes de collecte de données, à étoffer les connaissances quant à la comparabilité des deux systèmes de données sur la santé et à fournir un modèle pour de futures comparaisons internationales. Plusieurs études réalisées antérieurement avaient pour but d'étudier les comparaisons transculturelles des caractéristiques sociales, économiques et politiques au moyen d'ensembles de données distincts, mais elles présentent plusieurs limites méthodologiques. Dans une étude comparant les tendances de l'immigration au Canada et aux États-Unis, Pryor and Long¹ ont postulé que la gamme de données disponibles, les univers inclus dans l'enquête, la qualité des sources de données et la diversité des concepts et des définitions sont des éléments qui entravent la comparabilité. Dans une étude des condamnations au criminel et des sanctions en Europe, Barclay² a constaté que les écarts importants entre les niveaux et degrés des sanctions étaient dus en partie à des différences de pratiques d'enregistrement. Lors de l'analyse de données provenant de trois enquêtes transculturelles se distinguant par la taille et la gamme des régions géographiques, Dow³ a observé que les effets de plan de sondage sur les ensembles de données donnent lieu à une augmentation de la variance et du risque d'erreur de première espèce. Enfin, selon Scheuch⁴, les problèmes que posent la recherche et les comparaisons transculturelles au moyen d'ensembles de données différents ont trait à la signification des questions, à l'équivalence des indicateurs, à l'unité d'analyse et aux problèmes administratifs propres à chaque étude.

¹ Pryor, Edward T. and John F. Long (1987). "The Canada-United States Joint Immigration Study: Issues in Data Comparability." *Workshop on International Migration Data*, Ottawa, Canada.

² Barclay, Gordon C. (2000). "The Comparability of Data On Convictions and Sanctions: Are International Comparisons Possible?" *European Journal On Criminal Policy and Research* 8: 13-26.

³ Dow, Malcolm M. (1993). "Saving the Theory: On Chi-Square Tests With Cross-Cultural Survey Data." *Cross-Cultural Research* 8:3,4: 247-276.

⁴ Scheuch, Erwin K. (1993). "The Cross-Cultural Use of Sample Surveys: Problems of Comparability." *Historical Social Research* 18: 104-138.

Les résultats d'enquête peuvent diverger même si l'on tient compte de la méthodologie propre à chaque ensemble de données. Lors d'une étude des données sur la garde d'enfants provenant des enquêtes réalisées aux États-Unis, Raley et al.⁵ ont constaté que, si l'on tient compte des effets particuliers à chaque étude, les différences méthodologiques (période de référence de l'étude, diversité des méthodes de triage des répondants) continuent de produire des écarts importants dans les résultats. Bien que les débats de ce genre sur la validité des études comparatives se poursuivent, l'ECCES permet d'étoffer les données déjà publiées sur le sujet et produit un ensemble de données entièrement comparables sur le Canada et les États-Unis.

⁵ Raley, et al. (2000). "The Quality and Comparability of Child Care Data in United States Surveys." *Social Science Research* 29: 356-381.

3. Objectifs

Les principaux objectifs de l'ECCES sont :

- élaborer, mettre en œuvre et documenter une collaboration entre bureaux nationaux de la statistique en vue de réaliser des enquêtes conjointes sur la santé auprès de leur population nationale respective;
- s'appuyer sur les connaissances acquises durant la réalisation de l'ECCES pour modifier ou perfectionner les questionnaires des enquêtes nationales sur la santé en cours dans les deux pays, afin d'accroître la comparabilité des données provenant de ces enquêtes; et
- produire un ensemble de données hautement comparables sur les populations du Canada et des États-Unis que les chercheurs pourront utiliser pour étudier les effets des changements dans les systèmes de santé, les soins de santé, l'état de santé et l'état fonctionnel, et qui permettront de réaliser des études sur la méthodologie des enquêtes.

4. Contenu de l'enquête

La présente section donne un résumé général du contenu sélectionné pour l'ECCES. Ce contenu est fondé sur certaines parties des questionnaires utilisés par les deux pays pour réaliser leurs enquêtes sur la santé régulières, à savoir la « National Health Interview Survey (NHIS) » aux États-Unis, et l'Enquête sur la santé dans les collectivités canadiennes (ESCC) ainsi que l'Enquête nationale sur la santé de la population (ENSP) au Canada. Les sujets convertis par l'ECCES sont résumés au tableau 4.1.

Tableau 4.1 Modules du questionnaire

| | |
|--|---|
| État de santé général | Limitation des activités aux États-Unis |
| Restriction des activités (module aléatoire provenant du Canada) | (module aléatoire provenant des États-Unis) |
| Problèmes de santé chroniques | Test PAP |
| Dépression | Mammographie |
| Utilisation des services – Santé mentale | Visites chez le dentiste |
| Usage du tabac | Assurance |
| État de santé – Indice de l'état de santé (module aléatoire provenant du Canada) | Limitation des activités professionnelles (module aléatoire provenant des États-Unis) |
| Taille et poids | Satisfaction des patients |
| Utilisation des soins de santé | Activités physiques |
| Consommation de médicaments | Renseignements sociodémographiques |

Les enquêtes sur la santé réalisées à l'heure actuelle aux États-Unis et au Canada sont assez semblables en ce qui concerne le contenu; autrement dit, elles couvrent la même gamme de questions relatives à la santé, comme les problèmes de santé chroniques et aigus, la santé mentale et physique, la couverture des services médicaux et des soins de santé par une assurance, etc. Elles diffèrent principalement en ce qui concerne la mesure de certaines entités ayant trait à la santé. Autrement dit, la conceptualisation de certaines entités est différente et, par conséquent, les questions posées au sujet de ces entités sont également différentes. Il en est particulièrement ainsi de ce que nous pourrions appeler, de façon générale, l'« état fonctionnel ».

Chaque enquête suit une approche différente pour mesurer l'état fonctionnel. Les enquêtes canadiennes s'appuient sur un ensemble de questions couvrant la « limitation d'activités » et l'« indice de l'état de santé » (Health Utility Index) qui a été élaboré à l'Université McMaster au Canada. La NHIS comporte à la fois une mesure de limitation dans les principales activités et un ensemble de questions de limitation fonctionnelle. Toutes ces mesures sont manifestement apparentées, tant à première vue que d'après les analyses empiriques, mais elles produisent des estimations différentes des niveaux et des répartitions de l'état fonctionnel dans les populations. Ces mesures sont incluses toutes les quatre dans l'ECCES afin de pouvoir examiner ces différences et de donner aux chercheurs l'occasion d'étudier les propriétés de chacune de ces mesures sur deux populations. Les quatre modules de l'état fonctionnel sont assortis de l'annotation « module aléatoire » (tableau 4.1) pour indiquer que tous les répondants ont répondu aux quatre modules, mais dans un ordre aléatoire. La randomisation a été utilisée pour essayer de

réduire le biais de répondant qu'aurait pu causer la fatigue ou l'impatience des répondants tenus de répondre à des questions qu'ils pourraient avoir jugées répétitives.

Les questions sur l'état fonctionnel et la plupart des autres questions ont été posées aux répondants canadiens ainsi qu'américains. Cependant, à cause de différences dans les exigences en matière de procédures et de confidentialité, ainsi qu'entre les deux pays proprement dit, certains groupes de questions ont été posées uniquement aux répondants canadiens ou américains et d'autres ne peuvent être comparées directement sans utiliser de facteur d'ajustement. Ces différences sont les suivantes :

- A. Des ensembles distincts de questions sur les antécédents raciaux ou culturels ont été posées aux répondants du Canada et des États-Unis.
- B. À cause de différences entre les systèmes de santé, diverses questions sur la couverture de l'assurance-maladie ont été posées aux répondants des États-Unis uniquement.
- C. Les comparaisons transnationales du revenu ne peuvent être faites sans apporter une correction pour les différences de devises entre les deux pays. Pour réaliser des comparaisons du revenu, deux taux de change sont fournis avec les données. Le premier permet de convertir les dollars canadiens en dollars américains et le second, les dollars américains en dollars canadiens. Comme les taux de change fluctuent quotidiennement, les valeurs fournies avec les données sont les médianes des taux de change quotidiens en vigueur durant la collecte des données. Pour contourner les problèmes de conversion de devises, les quintiles de revenu sont également inclus dans les données. Ces variables indiquent à quel endroit dans la répartition relative du revenu de chaque pays le revenu déclaré d'un répondant ou d'un ménage se situe et elles sont directement comparables entre pays sans aucun ajustement. Veuillez consulter la documentation sur les variables dérivées et groupées pour plus de précisions sur ces variables.

5. Plan d'échantillonnage

5.1 Population cible

L'ECCES vise les membres de 18 ans et plus des ménages du Canada et des États-Unis. Sont exclues du champ de l'enquête les personnes résidant en établissement de soins de santé, ainsi que les détenus des prisons et les membres à temps plein des Forces canadiennes ou américaines. Au Canada, les trois territoires du Nord (Yukon, Territoires du Nord-Ouest et Nunavut) sont également exclus. Pareillement, aux États-Unis, les territoires des États-Unis (Porto Rico, les Île vierges américaines, les Samoa américaines, Guam et le Commonwealth des îles Marianne du Nord) sont exclus du champ de l'enquête, mais les résidents du district de Columbia y sont inclus.

5.2 Plan d'échantillonnage et répartition de l'échantillon

L'échantillon de l'ECCES a été conçu de façon à produire des estimations fiables à l'échelle nationale pour trois groupes d'âge (18 à 44 ans, 45 à 64 ans et 65 ans et plus) selon le sexe. Statistique Canada et le NCHS étaient responsables, respectivement, de la conception de l'échantillon pour leur pays. Afin de produire des estimations nationales fiables pour les trois groupes d'âge selon le sexe et de respecter le budget accordé pour l'enquête, il a été établi que celle-ci devrait être réalisée auprès d'un échantillon de 3 500 répondants au Canada et de 5 000 répondants aux États-Unis. Ces tailles d'échantillon ont été augmentées avant la collecte des données pour tenir compte des unités hors du champ de l'enquête et des cas prévus de non-réponse.

L'échantillon de l'ECCES a été stratifié selon la province au Canada et selon quatre régions géographiques aux États-Unis (Nord-Est, Midwest, Ouest et Sud). Dans chaque pays, l'échantillon a été réparti entre les strates proportionnellement à la taille de la population de ces dernières.

5.3 Échantillonnage des ménages

Plusieurs approches ont été envisagées pour sélectionner l'échantillon, mais la nécessité d'appliquer la même méthode dans les deux pays limitait les options. La méthode de sélection de l'échantillon offrant la meilleure comparabilité entre les deux pays était la méthode de composition aléatoire (CA). Chaque organisme était responsable de tirer son propre échantillon.

5.3.1 Base de sondage au Canada

L'échantillonnage des ménages par composition aléatoire (CA) repose sur la méthode d'élimination des banques non valides (EBNV)⁶. Au départ, la base de sondage consiste en une liste de toutes les « banques de 100 numéros » possibles, chacune contenant 100 numéros de téléphone consécutifs ayant en commun les huit premiers de leurs dix chiffres. Par exemple, les numéros allant de 617-555-7100 à 617-555-7199 constituent une banque de 100 numéros de téléphone qui ont en commun les huit premiers chiffres. En utilisant tous les codes régionaux actifs au Canada et tous les préfixes actifs à l'intérieur de ces codes régionaux, on obtient une base de sondage contenant toutes les « banques de 100 numéros » possibles. Puis, on catégorise comme étant une banque non valide toute banque qui ne contient aucun numéro de téléphone résidentiel et on élimine ces banques non valides de la base de sondage à mesure qu'elles sont repérées. Il convient de souligner que ces banques ne sont éliminées que lorsqu'on possède des preuves provenant de diverses sources qu'elles ne sont pas valides. Toute banque pour laquelle on ne dispose d'aucun renseignement est retenue dans la base de sondage. Les banques ainsi retenues sont groupées pour former des strates de composition aléatoire (CA). Dans chaque strate CA, on sélectionne une banque au hasard (p. ex., 617-555-71XX) et l'on génère de façon aléatoire les deux derniers chiffres du numéro de téléphone (c.-à-d. un nombre compris entre 00 et 99) pour créer un numéro de téléphone complet à dix chiffres. On répète cette procédure jusqu'à ce qu'on ait produit le nombre requis de numéros de téléphone dans la strate CA. Souvent, le numéro généré n'est pas en service ou est hors du champ de l'enquête et, par conséquent, de nombreux numéros supplémentaires doivent être générés pour atteindre la taille visée de l'échantillon. Ce taux de réussite, appelé *taux de succès*, varie selon la région. Pour l'ECCES, au Canada, le taux de succès varie de 43 % à 58 % à l'échelle provinciale.

5.3.2 Base de sondage aux États-Unis

Aux États-Unis, l'échantillon de l'ECCES a été tiré à partir d'une base de sondage par composition aléatoire (CA) assistée par liste⁷. La méthode assistée par liste s'appuie sur des combinaisons par préfixe des codes régionaux ou des codes de bureau central pour créer une base de sondage comprenant des banques de 100 numéros de téléphone consécutifs (p. ex., 301-937-5900 à 301-937-5999). On a sélectionné l'échantillon par composition aléatoire au moyen du système d'échantillonnage GENESYS (système propriétaire de Donnelley Marketing Information Systems (DMIS)). Après avoir créé ces banques de 100 numéros, GENESYS repère celles dont la probabilité de contenir des numéros résidentiels actifs est très faible. Ces numéros sont éliminés de la base de sondage. À partir des banques de 100 numéros retenues, numéros que l'on nomme numéros de téléphone résidentiels publiés dans l'annuaire 1+, on sélectionne un échantillon aléatoire de numéros de téléphone complets à dix chiffres de telle façon qu'on connaisse la probabilité de sélection et qu'elle soit la même pour tous les numéros.

⁶ Norris, D.A. and Paton, D.G. (1991). *Canada's General Social Survey: Five Years of Experience*. Survey Methodology, 17, 227-240.

⁷ Lepkowski, J.M. (1988). *Telephone Sampling Methods in the United States*. In Telephone Survey Methodology (Eds. R. Groves et al.), 73-98. New York: John Wiley and Sons.

Le système GENESYS est doté d'une fonction permettant l'élimination par triage d'une partie des numéros non valides à titre d'activité préliminaire de préparation de l'échantillon. Le triage est réalisé par DMIS en trois phases. À la première phase, l'échantillon est apparié à une liste de numéros de téléphone résidentiels publiés dans l'annuaire et tout numéro apparié est retenu pour la troisième phase. Puis, la partie restante de l'échantillon est également appariée à une liste de numéros commerciaux provenant des pages jaunes, afin d'exclure les numéros commerciaux de l'échantillon. À la deuxième phase, un système appelé GENESYS-ID trie l'échantillon restant pour éliminer une partie des numéros non valides. Ce système, exploité sur des ordinateurs personnels dotés d'un matériel et de logiciels spéciaux compose automatiquement tous les numéros de téléphone pour déceler les signaux d'interception indiquant que le numéro n'est pas valide. À la troisième et dernière phase, le fichier des numéros de téléphone retenus (après avoir éliminé les numéros commerciaux et les numéros non valides) est fusionné au fichier des numéros résidentiels publiés dans l'annuaire qui ont été retenus à la première phase. Les numéros résultant de cette troisième phase sont envoyés au système d'Interview téléphonique assistée par ordinateur (IITAO).

5.4 Sélection des répondants

Comme on l'a mentionné plus haut, l'objectif de l'ECES consiste à produire des estimations fiables à l'échelle nationale pour six domaines, à savoir trois groupes d'âge (18 à 44 ans, 45 à 64 ans et 65 ans et plus) selon le sexe. Par la méthode de composition aléatoire (CA), il est difficile de contrôler la composition de l'échantillon, puisque l'âge et le sexe des répondants sont inconnus au départ. Comme les hommes de 65 ans et plus ne représentent que 7 % de la population et que 13 % seulement des ménages comptent au moins un homme de 65 ans et plus, un tirage purement aléatoire des répondants parmi les membres adultes des ménages aurait nécessité un échantillon de très grande taille pour garantir des estimations fiables pour ce groupe. Dans le contexte de l'ECES, le groupe des 65 ans et plus est important. Pour éviter de devoir sélectionner un échantillon excessivement grand et pour respecter les contraintes opérationnelles et budgétaires, il a été décidé d'augmenter la probabilité de sélection des personnes de 65 ans et plus.

À cette fin, l'application informatique a été conçue de façon à sélectionner aléatoirement le répondant uniquement parmi les membres de 65 ans et plus du ménage quand au moins un membre du ménage faisait partie de ce groupe. Pour les ménages ne contenant que des personnes âgées de moins de 65 ans, le répondant a été sélectionné aléatoirement parmi tous les membres adultes du ménage. Cette stratégie a augmenté légèrement la représentation des personnes de 65 ans et plus dans l'échantillon, sans créer de distorsion excessivement importante comparativement à la répartition observée dans la population. Le principal inconvénient de cette approche est qu'elle exclut systématiquement de l'échantillon la population de moins de 65 ans vivant avec une ou plusieurs personnes de 65 ans et plus, ce qui pourrait introduire un biais dans l'échantillon si ces personnes ont des caractéristiques particulières. Par contre, elle évite d'obtenir des poids ayant une valeur extrême. On obtiendrait ce genre de poids pour la population de moins de 65 ans vivant avec une ou plusieurs personnes de 65 ans et plus si la probabilité de sélection diminuait au point d'être presque nulle. Pour cette raison, et pour s'assurer que la représentation des personnes de 65 ans et plus soit suffisante, il a été conclu que le biais éventuel représentait un compromis acceptable.

Le tableau 5.1 donne une comparaison de la répartition non pondérée des répondants, d'une part, et de la population cible, d'autre part, pour les deux pays. Il montre aussi que les hommes, particulièrement ceux de 18 à 44 ans, sont sous-représentés parmi les répondants. Le fait que les hommes, particulièrement ceux de ce groupe d'âge, soient plus difficiles à rejoindre et à interviewer est connu et n'est pas particulier à la présente enquête.

Tableau 5.1 Répartition non pondérée des répondants vs. répartition de la population cible au Canada et aux États-Unis (2002)

| Population | | Répondants aux États-Unis (%) | Population cible aux États-Unis (%) | Répondants au Canada (%) | Population cible au Canada (%) |
|-----------------------|---------------|--|--|--------------------------------|---|
| 18 à 44 ans | Hommes | 19.69 | 25.78 | 21.65 | 26.31 |
| | Femmes | 25.42 | 26.52 | 26.50 | 25.95 |
| 45 à 64 ans | Hommes | 14.92 | 15.35 | 14.64 | 15.90 |
| | Femmes | 17.68 | 16.34 | 15.92 | 16.21 |
| 65 ans et plus | Hommes | 8.26 | 6.85 | 9.64 | 6.85 |
| | Femmes | 14.03 | 9.16 | 11.81 | 8.78 |

Source : Recensement de la population de 1996 (1996) Statistique Canada et Current Population Survey (2002) U.S. Department of Commerce, Economics and Statistics Division, Bureau of the Census.

6. Collecte des données

6.1 Conception du questionnaire et méthode de collecte des données

L'autorisation légale de réaliser l'ECCES a été obtenue aux termes de protocoles distincts et les opérations d'interview des répondants ont été exécutées sous l'auspice d'autorités juridiques distinctes dans les deux pays.

La collecte des données de l'ECCES auprès des répondants des États-Unis a été autorisée en vertu de l'article 306 de la *Public Health Service Act*. Cette loi a aussi procuré la base légale (section 308 (d)) pour assurer la confidentialité des données. Les données sont protégées contre toute publication qui pourrait compromettre la vie privée des répondants, en vertu de la *Privacy Act of 1974* et de ses mises à jour. La conception de l'ECCES et les questionnaires ont été révisés par la *NCHS Institutional Review Board* afin de s'assurer que les droits des répondants étaient protégés. Finalement, l'*Office of Management and Budget* révisé et approuve les enquêtes afin d'assurer le public que le gouvernement ne leur impose pas un fardeau indu principalement en confirmant que l'effort de recherche n'est pas une répétition d'autres efforts déployés par le gouvernement.

Faisant partie du *United States Institutional Review Process*, les employés de Statistique Canada travaillant sur l'enquête ont dû être assermentés. À titre d'agent du NCHS, les employés de Statistique Canada affectés à l'enquête étaient tenus légalement responsables de respecter les politiques et les procédures du NCHS relatives à la confidentialité lors de la collecte des données auprès des répondants des États-Unis.

Au Canada, Statistique Canada a procédé à la collecte des données de l'enquête aux termes de la *Loi sur la statistique*, qui donne à l'Agence le mandat très étendu consistant à recueillir, compiler, analyser, dépouiller et publier des renseignements statistiques sur les conditions économiques, sociales et générales du pays et de la population, et de produire des statistiques concernant une liste fort détaillée de questions particulières, dont la santé. La composante canadienne de l'ECCES a été autorisée et prescrite en bonne et due forme aux termes de l'article 7 de la *Loi sur la statistique*. En outre, de par la loi, les renseignements recueillis en vertu de la *Loi sur la statistique* sur des personnes, des entreprises ou des organismes qui peuvent être identifiés ne peuvent être diffusés à qui que ce soit en dehors de Statistique Canada sans le consentement de ces personnes, entreprises ou organismes, et ne peuvent être consultés en vertu d'aucune autre loi, telle que la *Loi sur l'accès à l'information*. L'obligation d'assurer le respect de la confidentialité des renseignements recueillis aux termes de la *Loi sur la statistique* incombe aux employés de Statistique Canada de prêter serment de discrétion et s'exposent à des peines allant de l'amende à l'emprisonnement pour toute violation du secret professionnel.

Les questions de l'ECCES ont été posées par la méthode d'interview téléphonique assistée par ordinateur (ITAO), qui offre un certain nombre d'avantages en ce qui concerne la qualité des données par rapport à d'autres méthodes de collecte. Premièrement, le texte des questions, y compris les périodes de référence et les pronoms, sont personnalisés automatiquement d'après des facteurs comme l'âge et le sexe du répondant, la date de l'interview et les réponses aux

questions précédentes. Deuxièmement, des règles de vérification sont appliquées automatiquement pour repérer les réponses incohérentes ou non incluses dans la fourchette de valeurs permises, et des messages-guides apparaissent à l'écran quand une entrée invalide est enregistrée. Le répondant reçoit une rétroaction immédiate et l'intervieweur peut corriger toute incohérence. Troisièmement, l'application saute automatiquement les questions qui ne sont pas applicables à la personne interviewée.

6.1.1 Traduction

Au Canada, la Politique de communication du gouvernement du Canada et la *Charte canadienne des droits et libertés* établissent le principe de l'égalité du français et de l'anglais comme langues officielles du Canada. La charte confère au public canadien le droit de communiquer avec le gouvernement du Canada dans l'une ou l'autre langue. Les communications avec le public doivent se faire dans les deux langues et les services offerts au public doivent aussi l'être dans les deux langues, comme l'exige la *Loi sur les langues officielles*. Pour respecter ces lignes directrices, tous les questionnaires d'enquête et la documentation connexe sont élaborés, produits et diffusés simultanément dans les deux langues officielles, à savoir l'anglais et le français, et les deux versions sont équivalentes.

Aux États-Unis, selon les données du United States Bureau of the Census⁸, plus de 15 % de la population à domicile des États-Unis parle une autre langue que l'anglais. Afin d'obtenir des taux de réponse élevés, il est essentiel que les ménages ne parlant pas l'anglais soient inclus dans l'échantillon dans la mesure du possible. Pour résoudre ce problème, les modules de l'ECCES ont été traduits en espagnol à l'intention des résidents des États-Unis et des méthodes appropriées ont été mises au point (écrans d'aide au répondant appropriés, questions d'approfondissement particulières, intervieweurs parlant l'espagnol) pour mener les interviews auprès des ménages parlant l'espagnol. Tous les employés affectés à l'interview des ménages de langue espagnole étaient bilingues.

Pour la version espagnole du questionnaire, le NCHS a confié la traduction en sous-traitance à une entreprise externe. Une réunion d'examen de la version espagnole a eu lieu dans les locaux du NCHS au début de juin 2002. Les traducteurs n'étaient pas agréés, mais possédaient une grande expérience de la traduction de questionnaires d'enquête et avaient travaillé avec le NCHS à plusieurs occasions.

6.2 Interview

La collecte des données s'est déroulée du 4 novembre 2002 au 31 mars 2003. Des opérations de collecte supplémentaires ont eu lieu pendant plusieurs semaines en avril et en juin 2003, uniquement pour l'échantillon des États-Unis, afin d'essayer de convaincre de participer à l'enquête certaines personnes qui avaient refusé antérieurement de le faire. Dans chaque ménage

⁸ United States Department of Commerce, Economics and Statistics Division, Bureau of the Census, 2000.

sélectionné, on a demandé à une personne bien informée de 18 ans et plus de fournir des renseignements démographiques de base sur tous les membres du ménage. Puis, un membre du ménage de 18 ans ou plus a été sélectionné au hasard pour participer à une interview plus approfondie.

Dans les cas où le répondant sélectionné au hasard était incapable de participer à l'interview, un autre membre bien informé du ménage a fourni les renseignements à son sujet. Cette méthode porte le nom d'interview par procuration. Les personnes qui ont répondu par procuration ont pu fournir des réponses exactes à la plupart des questions, mais dans certains cas, elles n'étaient pas suffisamment bien informées pour répondre aux questions plus délicates ou plus personnelles, de sorte que certaines questions posées par procuration peuvent demeurer sans réponse. Par conséquent, on s'est efforcé par tous les moyens de réduire au minimum le nombre d'interviews par procuration.

Les interviews canadiennes et américaines ont été réalisées par les employés permanents des bureaux régionaux de Statistique Canada au moyen du même questionnaire. Les intervieweurs sont des employés recrutés et formés spécialement pour conduire des enquêtes par la méthode d'interview assistée par ordinateur et la plupart sont des intervieweurs chevronnés. Tous les intervieweurs ont participé à une séance de formation et reçu un manuel à titre d'outil de référence. Les interviews ont été réalisées en trois langues, à savoir en français et en anglais au Canada, et en espagnol et en anglais aux États-Unis. La durée de l'interview était d'environ 30 minutes.

6.3 Réduction de la non-réponse

Avant la première prise de contact par l'intervieweur, une lettre d'introduction a été envoyée par la poste à chaque logement sélectionné pour laquelle une adresse postale valide était disponible. Cette lettre expliquait aux répondants l'importance de l'enquête et les assurait du respect de la confidentialité des renseignements qu'ils fourniraient.

Le contenu des lettres d'introduction était presque identique pour les deux pays, les différences tenant à la mention de la loi en vertu de laquelle était réalisée l'enquête (la *Loi sur la statistique* pour le Canada et la *Public Health Service Act* pour les États-Unis) et aux organismes concernés. Les lettres ont été rédigées de façon à satisfaire les critères institutionnels des deux organismes, en reflétant les efforts déployés par leurs employés respectifs en vue de rendre la lettre aussi concise et intelligible que possible au niveau de la huitième année d'études. Statistique Canada s'est chargé d'envoyer par la poste les lettres d'introduction à l'intention des membres de l'échantillon du Canada, et le NCHS, celles destinées aux membres de l'échantillon des États-Unis, par la voie du centre d'envoi par la poste du United States Public Health Service établi à Rockville, au Maryland.

Les intervieweurs ont reçu l'instruction de faire tous les efforts raisonnables pour réaliser les interviews. Si le moment de l'appel de l'intervieweur était mal choisi, ce dernier prenait rendez-vous pour un autre appel à un moment plus opportun. En cas d'absence au moment de l'appel, de nombreuses autres tentatives ont été faites pour prendre contact avec le ménage. Aux

personnes faisant partie de l'échantillon qui ont refusé au départ de participer à l'enquête, le bureau régional a envoyé une lettre pour souligner l'importance de l'enquête et de la coopération du ménage. Ensuite, un intervieweur principal, un superviseur de projet ou un autre intervieweur a fait un autre appel pour essayer de convaincre le répondant sélectionné de l'importance de sa participation à l'enquête. Durant les derniers mois de la collecte des données, les efforts se sont concentrés sur les cas de non-réponse et sur certaines personnes qui avaient antérieurement refusé de participer à l'enquête.

7. Traitement des données

7.1 Vérification

La vérification des données est effectuée en grande partie au moment de l'interview par l'application d'interview téléphonique assistée par ordinateur (ITAO). Les intervieweurs ne peuvent entrer des valeurs situées en dehors des fourchettes permises et l'enchaînement des questions est contrôlé par les instructions « passer à » programmées. Par exemple, l'application d'ITAO assure que toute question ne s'appliquant pas au répondant ne soit pas posée. La valeur des zones correspondant à ces questions est fixée automatiquement à « non applicable », ce qui signifie que la question n'a pas été posée au répondant parce qu'il ne faisait pas partie de la catégorie d'intérêt (en fonction de l'âge, du sexe, de l'état matrimonial, etc.). Les questions sur le travail ne sont pas posées aux jeunes et les questions sur le cancer de la prostate ne sont pas posées aux femmes.

La valeur « non applicable » ne doit pas être confondue avec la valeur « non déclaré ». Cette dernière comprend les situations où le répondant ne se souvient pas de la réponse et ne peut la donner, mais où il reste au téléphone et continue de répondre au questionnaire. Par exemple, si on demande une date particulière, le répondant pourrait se souvenir de l'année, mais non du mois et du jour. Le code « non déclaré » sera inscrit dans les zones réservées au mois et au jour. En outre, si la personne raccroche après avoir répondu à quelques questions, le code « non déclaré » sera attribué à toutes les questions restantes auxquelles elle n'aura pas répondu. Une personne peut aussi refuser de répondre à une question ou ne pas connaître la réponse à une question particulière, auquel cas le code « refus » ou « ne sait pas » est attribué.

En réponse à certains types de déclarations incohérentes ou inhabituelles, l'application affiche un message d'avertissement, mais aucune mesure correctrice n'est prise au moment de l'interview. Au besoin, on a élaboré des règles de vérification qui ont été appliquées après la collecte des données, au Bureau central de Statistique Canada. En général, les incohérences sont corrigées en attribuant la valeur « non déclaré » à l'une des variables en cause, ou aux deux.

7.2 Codage

Des catégories de réponses précodées ont été fournies pour toutes les variables appropriées.

Il est possible de donner une réponse ouverte à plusieurs questions de l'ECCES. Pour certaines de ces questions, on a attribué aux réponses le code d'une catégorie énumérée si les renseignements faisaient double emploi avec une catégorie existante.

7.3 Création de variables dérivées et groupées

Pour faciliter l'analyse des données, plusieurs variables du fichier ont été dérivées d'après les items du questionnaire de l'ECCES. Ces variables sont appelées « variables dérivées » au Canada et « variable recodées » aux États-Unis. Le cinquième caractère du nom des variables dérivées est habituellement un D ou un G. Dans certains cas, les variables dérivées sont simplement obtenues en regroupant des catégories de réponse. Dans d'autres, plusieurs variables sont combinées pour en former une nouvelle. Les détails du calcul de ces variables plus complexes figurent dans le document sur les variables dérivées.

7.4 Pondération

Le principe sur lequel s'appuie l'estimation dans le cas d'un échantillon probabiliste comme celui de l'ECCES veut que, outre elle-même, chaque personne faisant partie de l'échantillon en représente plusieurs autres qui ne font pas partie de l'échantillon. Par exemple, dans un échantillon aléatoire simple où chaque personne a une probabilité de 1/50 (ou 2 %) d'être choisie, chaque personne en représente 50. Conformément à la terminologie utilisée ici, chaque personne a un facteur de pondération dans la population (ou un poids) de 50.

La phase de pondération est l'étape où l'on calcule le poids appliqué à chaque personne échantillonnée. Le poids figure dans le fichier de microdonnées et doit être utilisé pour produire des estimations significatives d'après les résultats de l'enquête. Par exemple, si l'on veut estimer le nombre de personnes qui fumaient, on choisit les enregistrements se rapportant aux personnes qui possèdent cette caractéristique et on additionne les poids inscrits sur ces enregistrements. La méthode employée pour calculer ces poids est décrite en détail à la section 8.

7.5 Suppression de renseignements confidentiels

Il convient de souligner que le fichier de microdonnées à grande diffusion décrit ici diffère en plusieurs points importants du « fichier maître » de l'enquête gardé par les organismes. Ces différences sont dues aux mesures prises pour protéger l'anonymat des répondants à l'enquête. La confidentialité des renseignements fournis par les répondants est assurée grâce à la suppression de valeurs individuelles, au groupement de variables et au plafonnement des valeurs des variables dans le fichier de microdonnées à grande diffusion. Veuillez consulter le tableau 7.1 pour la liste complète des variables qui ont été regroupées ou dont la valeur a été plafonnée dans le fichier. Le Dictionnaire de données et la Documentation sur les variables dérivées et groupées fournissent aussi les définitions de toutes ces variables dérivées.

Tableau 7.1 Variables regroupées ou dont la valeur est plafonnée

| Nom de la variable | Définition | Groupement ou plafonnement |
|---------------------------|---|--|
| DHJ1GNHH | Nombre de personnes dans le ménage | Plafonnée à 5 personnes ou plus |
| DHJ1GAGE | Âge | Plafonnée à 85 ans et plus |
| HUJ1GDEX | Troubles de dextérité - code fonctionnel | Regroupée en 3 catégories |
| HUJ1GHER | Troubles de l'ouïe - code fonctionnel | Regroupée en 3 catégories |
| HUJ1GMOB | Troubles de mobilité - code fonctionnel | Regroupée en 4 catégories |
| HUJ1GSPE | Troubles d'élocution - code fonctionnel | Regroupée en 2 catégories |
| HUJ1GVIS | Troubles de la vision - code fonctionnel | Regroupée en 5 catégories |
| HCJ1G2A | Nombre de consultations - médecin de famille/omnipraticien | Plafonnée à 31 ou plus |
| HCJ1G2B | Nombre de consultations - spécialiste de la vue | Plafonnée à 12 ou plus |
| HCJ1G2C | Nombre de consultations - chiropraticien | Plafonnée à 31 ou plus |
| HCJ1G2D | Nombre de consultations - infirmière | Plafonnée à 12 ou plus |
| HCJ1G2E | Nombre de consultations - dentiste | Plafonnée à 12 ou plus |
| HCJ1G2F | Nombre de consultations - physiothérapeute | Plafonnée à 31 ou plus |
| HCJ1G2G | Nombre de consultations - psychologue | Plafonnée à 12 ou plus |
| HCJ1G2H | Nombre de consultations - orthophoniste | Plafonnée à 12 ou plus |
| HCJ1G2I | Nombre de consultations - un autre médecin | Plafonnée à 12 ou plus |
| HCJ1GMC | Nombre de consultations avec des médecins | Plafonnée à 12 ou plus |
| CMJ1G01L | Santé mentale - nombre de fois a consulté professionnel de la santé | Plafonnée à 12 ou plus |
| HCJ1G01A | Nombre de nuits comme patient | Plafonnée à 31 ou plus |
| SDJ1GMS | État matrimonial | Regroupées en 4 catégories |
| SDJ1GHED | Plus haut niveau de scolarité postsecondaire atteint | Regroupées en 4 catégories |
| SDJ1GCBC | Pays de naissance - Canada seulement | Regroupées en 2 catégories |
| SDJ1GCBU | Pays de naissance - États-Unis seulement | Regroupées en 2 catégories |
| IWJ1GMSI | Revenu total du ménage - source principale | Regroupées en 5 catégories |
| IWJ1GTHI | Revenu total du ménage de toutes sources | Plafonnée à 30 000 \$ ou plus |
| IWJ1GTPI | Revenu personnel total de toutes sources | Plafonnée à 80 000 \$ ou plus |
| IWJ1GHEQ | Valeur nette de la propriété | Plafonnée à -500 000 \$ ou moins et à 500 000 \$ ou plus |

8. Pondération

Pour que les estimations produites à partir de données d'enquête soient représentatives de la population cible, et non pas seulement de l'échantillon proprement dit, l'utilisateur doit intégrer les poids de sondage dans le calcul des estimations. Un poids de sondage est attribué à chaque personne incluse dans l'échantillon final, c'est-à-dire l'échantillon de personnes qui ont répondu à l'enquête. Ce poids correspond au nombre de personnes de la population cible que représente le répondant.

Les poids de sondage ont été calculés séparément pour les échantillons canadien et américain, mais, dans les deux cas, selon la méthode décrite ci-après. Le tableau 8.1 donne un aperçu des divers ajustements qui font partie de la stratégie de pondération, dans l'ordre où ils ont été appliqués.

Tableau 8.1 Liste des ajustements de la pondération

| | |
|---|--|
| 0 | Poids CA initial |
| 1 | Élimination des numéros hors du champ de l'enquête |
| 2 | Ajustement pour la non-réponse des ménages |
| 3 | Ajustement pour les lignes téléphoniques multiples |
| 4 | Création du poids personne |
| 5 | Ajustement pour la non-réponse personne |
| 6 | Poststratification |

8.1 Ajustements appliqués au poids initial

Ajustement 0 – Poids CA initial

La méthode de sélection par composition aléatoire utilisée pour l'enquête produit un échantillon aléatoire simple stratifié (sans remise) de lignes téléphoniques résidentielles. Donc, un « poids CA initial » est donné par l'inverse de la probabilité de sélection d'une ligne téléphonique résidentielle à partir d'une liste de numéros de téléphone. Le poids CA initial (différent pour chaque strate) est donné par :

$$W_{initial} = \frac{\text{nombre total de numéros de téléphone dans la base de sondage } (N)}{\text{nombre total de numéros de téléphone échantillonnés aléatoirement à partir de cette base de sondage}}$$

Il convient de souligner qu'au Canada, certaines strates originales ont été regroupées pour des raisons de confidentialité.

Ajustement 1 – Élimination des numéros hors du champ de l'enquête

Les numéros de téléphone menant à des entreprises, des établissements, ainsi que les numéros hors-service sont des exemples de cas hors du champ de l'enquête en ce qui concerne une base de sondage téléphonique.

Le poids des numéros de téléphone échantillonnés qui sont hors du champ de l'enquête est fixé à 0 (au moyen d'un facteur d'ajustement fictif). Aux États-Unis, un certain nombre de numéros de téléphone sont demeurés des cas non résolus à la fin de la période de collecte des données (voir section 9.1). Tous les cas non résolus ont été retenus après le présent ajustement, mais leur poids de base a été multiplié par un facteur de correction donné par la proportion de numéro « dans le champ » de l'enquête ($P_{\text{dans le champ}}$) parmi tous les cas résolus (dans le champ et hors du champ de l'enquête). Le poids résultant a été obtenu en multipliant le poids initial par le facteur de correction suivant, au niveau de la strate :

$$A_1 = \begin{cases} 0 & \text{si hors du champ} \\ P_{\text{dans le champ}} & \text{non résolu (États - Unis seulement)} \\ 1 & \text{autrement} \end{cases}$$

Les enregistrements hors du champ de l'enquête (poids=0) ont alors été éliminés du fichier.

Ajustement 2 – Ajustement pour la non-réponse des ménages

Malgré tous les efforts déployés par les intervieweurs, un certain degré de non-réponse à l'échelle des ménages est inévitable. La non-réponse comprend toutes les situations suivantes : un refus, des circonstances spéciales, un obstacle linguistique, une absence temporaire ou à un problème informatique. La non-réponse des ménages est compensée en ajustant de façon proportionnelle le poids des ménages répondants. Cet ajustement, qui est effectué pour la strate, est donné par :

$$A_2 = \frac{\text{somme des poids de tous les ménages échantillonnés}}{\text{somme des poids des ménages répondants}}.$$

De nouveau, au Canada, certaines strates originales ont été regroupées pour des raisons de confidentialité.

Ajustement 3 – Ajustement pour les lignes téléphoniques multiples

Tout ménage comptant plusieurs lignes téléphoniques résidentielles énumérées dans la base de sondage CA a une plus grande probabilité d'être sélectionné dans l'échantillon qu'un ménage ne possédant qu'une seule de ces lignes. Donc, on applique aux ménages échantillonnés pour lesquels il est confirmé qu'ils possèdent plusieurs lignes téléphoniques un facteur de correction égal à l'inverse du nombre de lignes téléphoniques résidentielles dans le ménage. Il convient de souligner que cette information a été obtenue au début de l'interview. Pour réduire la variabilité des poids, il a été décidé de fixer à trois le nombre maximal de lignes téléphoniques pour cette

correction. Chaque ménage comptant plus de trois lignes téléphoniques a été considéré comme en ayant seulement trois. Cet ajustement multiplicatif, effectué pour la strate, est donné par :

$$A_3 = \frac{1}{\text{nombre de lignes téléphoniques vocales résidentielles dans le ménage (max 3)}} .$$

Nota : À cause d'un problème technique posé par l'application informatique, le nombre de lignes téléphoniques résidentielles dans le ménage n'est pas disponible pour les interviews réalisées durant les deux premières semaines de la période de collecte. Pour les cas où l'information n'est pas disponible, les poids ont été divisés par le nombre moyen de lignes téléphoniques résidentielles dans un ménage calculé d'après les enregistrements pour lesquels l'information était disponible. Cette correction a été effectuée séparément pour chaque taille de ménage (1, 2, 3, 4 ou 5 +), puisque le nombre de personnes dans le ménage influe sur le nombre de lignes téléphoniques.

Ajustement 4 – Création du poids personne

Cet ajustement convertit le poids du ménage en un poids personne. Comme il est mentionné plus haut, une seule personne de 18 ans et plus a été sélectionnée dans chaque ménage échantillonné. Pour réduire la variabilité des poids, il a été décidé de fixer à trois le nombre maximal de personnes de 18 ans et plus pour le présent ajustement pour l'échantillon des États-Unis. Chaque ménage échantillonné aux États-Unis comptant plus de trois personnes de 18 ans et plus a été considéré comme ne comptant que trois de ces personnes. L'ajustement multiplicatif nécessaire pour refléter cette sélection est donné par :

$$A_4 = \frac{1}{\text{probabilité de sélection dans le ménage}} .$$

Nota : Consulter la section 5 pour plus de précisions sur les probabilités inégales de sélection des personnes de 65 ans et plus.

Ajustement 5 – Ajustement pour la non-réponse personne

Cet ajustement consiste à compenser les effets de la non-réponse à l'échelle de la personne. Il arrive que, même si un ménage est considéré comme étant « répondant », l'information pour la personne sélectionnée dans le ménage ne soit pas complété. On considère les personnes sélectionnées pour lesquelles il en est ainsi comme des non-répondants et on ajuste la pondération des personnes sélectionnées répondantes appartenant à la même catégorie âge-sexe-strate-taille du ménage pour compenser. La taille du ménage (1, 2 ou 3 et plus) est utilisée parce qu'elle est corrélée au taux de réponse pour la personne sélectionnée. L'ajustement multiplicatif est donné par :

$$A_5 = \frac{\text{somme des poids de toutes les personnes sélectionnées dans une catégorie âge - sexe - strate - taille du ménage}}{\text{somme des poids des personnes sélectionnées répondantes dans une catégorie âge - sexe - strate - taille du ménage}}$$

Pour chaque sexe, les catégories d'âge utilisées sont 18 à 44 ans, 45 à 64 ans et 65 ans et plus. Au Canada, certaines strates originales ont été regroupées pour des raisons de confidentialité.

Ajustement 6 – Poststratification

Enfin, on a procédé à une poststratification pour s'assurer que la somme des poids finaux concorde avec les estimations de population pour certaines variables auxiliaires. Au Canada, les estimations de population sont fondées sur le Recensement de la population de 1996⁹ et aux États-Unis, sur la Current Population Survey d'octobre 2002¹⁰. Les variables auxiliaires utilisées pour créer les poststrates sont énumérées au tableau 8.2.

Tableau 8.2 Variables auxiliaires

| Canada | États-Unis |
|--|--|
| Âge (5 groupes): 18 à 34 ans 35 à 44 ans 45 à 54 ans 55 à 64 ans 65 ans et plus | Âge (5 groupes [*]) : 18 à 34 ans 35 à 44 ans 45 à 54 ans 55 à 64 ans 65 ans et plus |
| Sexe : Homme Femmes | Sexe : Homme Femme |
| Région : 1 – Atlantique 2 – Québec 3 – Ontario 4 – Prairies 5 – Colombie-Britannique | Race/groupe ethnique : 1 - Hispanique 2 - Non-Hispanique / Noir 3 - Non-Hispanique et non-Noir |

* Pour race/groupe ethnique = Hispanique ou non-Hispanique/Noir, trois groupes d'âge seulement ont été utilisés (18 à 44 ans, 45 à 64 ans et 65 ans et plus) à cause de contraintes de taille d'échantillon.

Pour les parties canadiennes et américaines de l'échantillon, cinq groupes d'âge (à une exception près aux États-Unis) ont été utilisés au lieu de trois afin d'obtenir une meilleure répartition de l'échantillon comparativement à la répartition par âge de la population. Cette mesure a également réduit l'effet des probabilités inégales de sélection des répondants (aucune personne de moins de 65 ans vivant avec une personne de plus de 65 ans n'a été sélectionnée, si bien que ce groupe était sous-représenté dans l'échantillon).

Cet ajustement tient également compte du fait qu'un faible pourcentage de ménages n'ont pas le téléphone. Comme les estimations de population donnent le nombre total de personnes dans la population cible, que le ménage reçoive un service téléphonique ou non, cet ajustement corrige en partie pour les adultes qui vivent dans un ménage n'ayant pas le téléphone. Il ne corrige pas

⁹ Recensement de la population de 1996. Statistique Canada. Ottawa.

¹⁰ Current Population Survey (2002). U.S. Department of Commerce, Economics and Statistics Division, Bureau of the Census.

les biais introduits si la non-réponse due au fait de ne pas avoir le téléphone n'est pas aléatoire.

L'ajustement multiplicatif est donné par :

$$A_6 = \frac{\textit{estimation de la population pour la classe de poststratification}}{\textit{somme des poids des personnes sélectionnées répondantes dans la classe de poststratification}} .$$

Poids final

Enfin, on calcule le poids final en multipliant le poids CA initial par les facteurs d'ajustement 1 à 6 :

$$W_{final} = W_{initial} \times A_1 \times A_2 \times A_3 \times A_4 \times A_5 \times A_6 .$$

Le poids final figure dans le fichier de données sous le nom de variable WT_SAM.

9 Qualité des données

9.1 Taux de réponse

Le taux de réponse global est de 65,5 % pour l'échantillon canadien et de 50,2 % pour l'échantillon américain. L'un des principaux problèmes que pose l'existence de numéros invalides dans l'échantillon CA est qu'il est difficile de déterminer si les numéros pour lesquels on n'a pas obtenu de réponse sont valides ou non. Au Canada, comme le nombre de compagnies de téléphone est faible, il est possible d'utiliser leurs listes pour valider les numéros. Aux États-Unis, le nombre plus grand de compagnies rend cette pratique impossible, si bien que la validité de plusieurs numéros demeure inconnue. Par conséquent, les taux de réponse pour le Canada et pour les États-Unis sont calculés en suivant des lignes directrices différentes.

9.1.1 Taux de réponse au Canada

Au total et après élimination des unités hors du champ de l'enquête, 5 355 ménages canadiens ont été sélectionnés pour participer à l'ECCES. Une réponse a été obtenue auprès de 3 858 de ces ménages sélectionnés, ce qui donne un taux de réponse global à l'échelle du ménage de 72,0 %. Parmi ces ménages répondants, 3 858 personnes (une par ménage) ont été sélectionnées pour participer à l'ECCES et 3 505 d'entre elles y ont effectivement participé, ce qui donne un taux de réponse à l'échelle de la personne de 90,9 %. Pour le Canada, ceci donne un taux de réponse global de **65,5 %**.

Taux de réponse au niveau du ménage (TRM)

$$\text{TRM} = \frac{\text{Nombre de ménages répondants}}{\text{Tous les ménages dans le champ de l'enquête}} .$$

Taux de réponse au niveau de la personne (TRP)

$$\text{TRP} = \frac{\text{Nombre de personnes sélectionnées répondantes.}}{\text{Toutes les personnes sélectionnées}}$$

$$\text{Taux de réponse global} = \text{TRM} \times \text{TRP} .$$

Voici un exemple du calcul du taux de réponse global pour le Canada.

$$\text{TRM} = 3\,858 / 5\,355 = 0,720 .$$

$$\text{TRP} = 3\,505 / 3\,858 = 0,909 .$$

$$\begin{aligned} \text{Taux de réponse global} &= 0,720 \times 0,909 \\ &= 0,655 \\ &= \mathbf{65,5 \%} . \end{aligned}$$

9.1.2 Taux de réponse aux États-Unis

Les taux de réponse aux États-Unis, fondés sur les lignes directrices du Council of American Survey Research Organizations (CASRO), ont été calculés conformément aux *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*¹¹ de l'American Association for Public Opinion Research (AAPOR) en utilisant les hypothèses de l'Association pour le taux de réponse n° 4.

Le taux de résolution cas aux États-Unis mesure la proportion de numéros de téléphone échantillonnés pour lesquels il a été possible de confirmer catégoriquement qu'il s'agissait d'un numéro résidentiel ou non résidentiel. Ce taux est de 80,4 % pour l'ECCES. L'appel de la majorité des numéros de téléphone dont le cas n'est pas résolu a abouti à une sonnerie sans réponse ou au contact avec une personne ou une machine qui a « raccroché » avant de s'identifier. (L'appel de la plupart des autres numéros dont le cas demeure non résolu a abouti à un répondeur ne précisant pas si l'appelant avait rejoint une résidence ou une entreprise commerciale.) Ce taux de résolution est l'une des composantes du taux de réponse global.

La deuxième composante du taux de réponse global est le taux de coopération, qui mesure la proportion de ménages connus auprès desquels une interview a été réalisée. Pour l'ECCES, le taux de coopération est de 62,4 %. Le taux de réponse CASRO non pondéré a ensuite été calculé en multipliant le taux de résolution (80,4 %) par le taux de coopération (62,4 %), ce qui donne un taux de réponse global aux États-Unis de 50,2 %. Des renseignements détaillés concernant la composition finale de l'échantillon et le taux de réponse aux États-Unis figurent au tableau 9.1.2.

Tableau 9.1.2 Composition finale de l'échantillon

| Catégorie | Fréquence |
|--|---------------|
| Nombre total de cas hors du champ de l'enquête | 17 437 |
| Nombre total de cas non résolus | 6 263 |
| Non-réponse (ménage connu ou supposé) | 3 117 |
| Interview réalisée (ménage connu) | 5 183 |
| Nombre total de numéros sélectionnés | 32 000 |

Taux de résolution (TR)

$$TR = \frac{N^{bre} \text{ de cas hors du champ} + N^{bre} \text{ de personnes non répondantes} + N^{bre} \text{ de personnes répondantes}}{N^{bre} \text{ total de numéros de téléphone sélectionnés}}$$

¹¹The American Association for Public Opinion Research (2004). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 3rd edition. Lenexa, Kansas: AAPOR.

Taux de coopération (TC)

$$TC = \frac{N^{bre} \text{ de personnes répondantes}}{N^{bre} \text{ de personnes non répondantes} + N^{bre} \text{ de personnes répondantes}} .$$

$$\text{Taux de réponse global} = TR \times TC .$$

Voici un exemple de calcul du taux de réponse global pour les États-Unis d'après l'information figurant au tableau 9.1.2.

$$TR = \frac{(17\,437 + 3\,117 + 5\,183)}{32\,000} = 0,804 .$$

$$TC = \frac{5\,183}{3\,117 + 5\,183} = 0,624 .$$

$$\begin{aligned} \text{Taux de réponse global} &= 0,804 \times 0,624 \\ &= 0,502 \\ &= \mathbf{50,2 \%} . \end{aligned}$$

9.2 Erreurs dans l'enquête

L'enquête permet de produire des estimations fondées sur les données recueillies auprès d'un échantillon de personnes. On aurait obtenu des estimations légèrement différentes si on avait procédé à un recensement complet en utilisant le même questionnaire, les mêmes intervieweurs, les mêmes superviseurs, les mêmes méthodes de traitement, etc. que ceux utilisés pour l'enquête. La différence entre les estimations tirées de l'échantillon et celles qui découlent d'un dénombrement complet effectué dans des conditions semblables s'appelle *erreur d'échantillonnage* de l'estimation.

Des erreurs qui ne sont pas liées à l'échantillonnage peuvent être commises à presque toutes les étapes d'une enquête. Il se peut que les intervieweurs comprennent mal les instructions, que les répondants donnent des réponses incorrectes, que les réponses soient mal enregistrées ou que des erreurs se produisent au moment du traitement et de la totalisation des données. Ces erreurs sont toutes des exemples d'*erreurs non dues à l'échantillonnage*.

9.2.1 Erreurs non dues à l'échantillonnage

Sur un grand nombre d'observations, les erreurs aléatoires ont peu d'effet sur les estimations tirées des données d'enquête. Toutefois, les erreurs qui se produisent systématiquement introduisent un biais dans ces estimations. Beaucoup de temps et d'effort ont été consacrés à la réduction des erreurs non dues à l'échantillonnage dans l'ECCES. Des mesures d'assurance de la qualité ont été appliquées à chaque étape du cycle de collecte et de traitement des données afin de contrôler la qualité des données. On a notamment fait appel à des intervieweurs hautement

qualifiés, donné à ceux-ci une formation poussée sur les méthodes d'enquête et sur le questionnaire, observé les intervieweurs afin de déceler les problèmes et testé l'application d'ITAO.

L'effet de la non-réponse sur les résultats de l'enquête constitue une source importante d'erreurs non dues à l'échantillonnage dans les enquêtes. L'ampleur de la non-réponse varie d'une non-réponse partielle (le fait de ne pas répondre à une ou à plusieurs questions) à une non-réponse totale. Dans le cas de l'ECCES, la non-réponse partielle a été minimale; une fois qu'ils ont commencé à répondre au questionnaire, les répondants ont eu tendance à aller jusqu'au bout. Il y a eu non-réponse totale lorsqu'un répondant a refusé de participer à l'enquête ou que l'intervieweur a été incapable de prendre contact avec le répondant sélectionné. On a traité les cas de non-réponse totale en corrigeant les poids des personnes qui ont répondu à l'enquête afin de compenser pour celles qui n'ont pas répondu. Des renseignements supplémentaires sur l'ajustement des poids pour la non-réponse figurent à la section 8.

9.2.2 Erreurs d'échantillonnage

Puisque les estimations d'une enquête par sondage comportent inévitablement des erreurs d'échantillonnage, de bonnes méthodes statistiques exigent que les chercheurs donnent aux utilisateurs une indication de la grandeur de cette erreur d'échantillonnage. La mesure de l'importance des erreurs dues à l'échantillonnage est fondée sur les écarts-types des estimations calculées d'après les données de l'enquête. Cependant, en raison de la diversité des estimations que l'on peut tirer d'une enquête, l'écart-type d'une estimation est habituellement exprimé en fonction de l'estimation à laquelle il se rapporte. La mesure résultante, appelée coefficient de variation (c.v.) de l'estimation, s'obtient en divisant l'écart-type de l'estimation par l'estimation proprement dite et s'exprime en pourcentage de l'estimation. Notons que le coefficient de variation est également appelé erreur-type relative.

Par exemple, supposons qu'on estime que 20 % des répondants de 18 ans sont des fumeurs et que l'écart-type de cette estimation est de 0,007. Alors, le coefficient de variation de l'estimation (exprimé en pourcentage) se calcule comme suit :

$$(0,007/0,20) \times 100 = 3,5 \% .$$

Statistique Canada et le NCHS utilisent couramment les coefficients de variation pour analyser les données et conseillent vivement aux utilisateurs qui produisent des estimations à partir du fichier de données de l'ECCES d'en faire de même. Pour des précisions sur l'analyse des données, consulter la section 10.

9.2.3 Estimation de la variance

Afin de déterminer la qualité d'une estimation et de calculer son c.v., on doit calculer l'écart-type de l'estimation. Cet écart-type est également nécessaire pour calculer les intervalles de confiance. Dans le cas de l'ECCES, il est recommandé d'estimer la variance par la méthode du bootstrap ou par la méthode du développement en série de Taylor.

9.2.3.1 Méthode du bootstrap pour l'estimation de la variance

Le plan de sondage de l'ECES étant complexe, on ne peut utiliser aucune formule simple pour estimer la variance. Par conséquent, il faut recourir à une méthode approximative, comme celle du bootstrap, qui permet de tenir compte de l'information contenue dans le plan de sondage lors du calcul des estimations de la variance. Cette méthode, qui peut être appliquée au moyen du programme Bootvar fourni avec les données et décrit à la sous-section suivante, est assez facile à utiliser.

La méthode du bootstrap appliquée aux données de l'ECES implique la sélection d'échantillons aléatoires simples appelés répétitions et le calcul de la variation des estimations d'une répétition à l'autre. Pour chaque répétition, on recalcule le poids de sondage de chaque enregistrement, puis on ajuste ces poids et on les poststratifie d'après les estimations de population connues, de la même façon que les poids initiaux, afin d'obtenir les poids bootstrap finaux.

Le processus complet (sélection d'échantillons aléatoires simples, recalcul et poststratification des poids pour chaque strate) est répété B fois, où B est grand. Pour l'ECES, on utilise B = 1 000 pour produire 1 000 ensembles de poids bootstrap, qui sont fournis avec le fichier de microdonnées à grande diffusion. Pour obtenir un estimateur bootstrap de la variance, on doit calculer l'estimation ponctuelle pour chacun des B échantillons. La variance de ces estimations est l'estimateur bootstrap de la variance. Le programme Bootvar a été développé pour permettre à l'utilisateur des données d'effectuer tous ces calculs.

9.2.3.1.1 Programme Bootvar pour l'estimation de la variance

Le programme Bootvar est disponible en format SAS ou SPSS. Il est constitué de macros qui calculent la variance de totaux, de rapports, de différences entre rapports et de régressions linéaires ou logistiques. Le programme Bootvar est fourni avec le fichier de microdonnées à grande diffusion, en même temps que les poids bootstrap et un document expliquant comment modifier et utiliser le programme pour l'adapter aux besoins de l'utilisateur.

9.2.3.2 Méthode du développement en série de Taylor (avec SUDAAN) pour l'estimation de la variance

On peut utiliser la méthode du développement en série de Taylor pour estimer la variance de totaux, de rapports, de régressions linéaires et de régressions logistiques. Dans le cas de l'ECES, on peut calculer les erreurs-types des estimations en utilisant un logiciel statistique tel que SUDAAN.¹² Aux fins de l'estimation de la variance, l'échantillon de l'ECES est traité comme un échantillon à deux degrés. Dans les applications SUDAAN, on utilise WTS_STR et SAMPLEID comme variables de strate dans l'énoncé SUDAAN NEST et WT_SAM est utilisé

¹² Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1997). *SUDAAN User's Manual, Release 7.5*. Research Triangle Park, NC: Research Triangle Institute.

dans l'énoncé WEIGHT. Notez que SAMPLEID doit d'abord être converti en format numérique (voir plus bas).

Les énoncés de plan de sondage avec SUDAAN suivants sont recommandés :

```
sampid = input(sampleid, 12.);
```

```
PROC SORT;  
  by WTS_STR SAMPID;
```

```
PROC...  DESIGN = WR;  
NEST    WTS_STR SAMPID;  
WEIGHT  WT_SAM;
```

Analyse de sous-ensembles de données. Fréquemment, les études portant sur des données d'enquête complexe sont limitées à certains sous-groupes de population, comme les personnes de 65 ans et plus. Certains utilisateurs suppriment tous les enregistrements hors du domaine d'intérêt (p. ex, personnes de moins de 65 ans) afin de travailler avec de plus petits fichiers de données et d'effectuer plus rapidement l'exécution des programmes. La procédure consistant à ne garder que certains enregistrements (et à supprimer d'après une liste d'autres enregistrements) porte le nom de définition de sous-ensembles de données. À partir d'une base de données dans laquelle sont définis des sous-ensembles pondérés comme il convient, on peut produire des estimations ponctuelles correctes (p. ex., estimations de moyennes de population de sous-groupes). Cependant, la plupart des progiciels qui analysent les données d'enquête complexes par application de la méthode de linéarisation de Taylor calculent incorrectement les erreurs-types des estimations produites pour des sous-ensembles de données. Souvent, la définition de sous-ensembles dans un ensemble de données complexes compromet la structure du plan d'échantillonnage, parce qu'on ne dispose plus de l'information complète sur ce dernier; la définition de sous-ensembles élimine d'importantes informations sur le plan de sondage nécessaires pour l'estimation de la variance. Il convient de souligner que SUDAAN possède une option SUBPOPN qui permet de cibler une sous-population tout en utilisant le fichier de données complet (sans définition de sous-ensembles) contenant toute l'information sur le plan de sondage. La présente section fournit deux stratégies pour calculer les variances au moyen de SUDAAN en tenant compte du plan d'enquête complexe. Pour plus d'information, consulter le manuel de SUDAAN.

Stratégie 1 Utiliser l'option MISSUNIT de l'énoncé NEST avec la méthode décrite plus haut pour la définition de sous-ensembles de données :

```
NEST          WTS_STR SAMPID / MISSUNIT;
```

Dans un plan d'échantillonnage avec remise comptant exactement deux UPE par strate, si certaines UPE sont éliminées de la base de sondage par suppression d'enregistrements hors population d'intérêt d'après une liste, l'option MISSUNIT de SUDAAN « fixe » l'estimation afin de produire des erreurs-types identiques à celles obtenues en utilisant un ensemble de données complet avec un énoncé SUBPOPN (voir la stratégie 2 plus bas). Il convient de souligner que d'autres calculs pour les effets de plan, le nombre de degrés de liberté et la standardisation doivent parfois être effectués différemment. Il incombe à l'utilisateur de vérifier l'exactitude des résultats qu'il obtient à partir de données pour lesquelles ont été définis des sous-ensembles.

Stratégie 2 Utiliser l'énoncé SUBPOPN avec la méthode décrite plus haut pour l'ensemble de données complet :

```
PROC ...      DESIGN = WR;
NEST          WTS_STR SAMPID;
WEIGHT        WT_SAM;
SUBGROUP      (variable names);
LEVELS...     ;
SUBPOPN       SPJ1_TYP = 1 and SEX = 2 / NAME="Analysis of Canadian
women";
```

Dans le présent exemple, l'utilisation de l'ensemble de données complet avec l'énoncé SUBPOPN limite l'analyse aux femmes canadiennes uniquement (SPJ1_TYP = 1 pour Canada et SEX = 2 pour femmes). L'utilisation de l'énoncé SUBPOPN équivaut à définir des sous-ensembles dans l'ensemble de données (c'est-à-dire enlever les cas américains), à cette différence près que les estimations résultantes de la variance sont fondées sur la structure complète du plan de sondage de l'ensemble de données.

10. Lignes directrices pour la totalisation, l'analyse et la diffusion

Cette section du guide décrit les lignes directrices que doivent suivre les utilisateurs qui totalisent, analysent, publient ou diffusent de quelque autre façon des données provenant du fichier de microdonnées de l'enquête. Ces lignes directrices devraient leur permettre de reproduire les chiffres déjà publiés par Statistique Canada et par le NCHS et de produire aussi des chiffres non encore publiés conformes aux lignes directrices établies.

10.1 Lignes directrices pour l'arrondissement

Afin que les estimations calculées d'après le fichier de microdonnées en vue d'être publiées ou diffusées de toute autre façon correspondent à celles produites par Statistique Canada et par le NCHS, il est vivement conseillé à l'utilisateur de les arrondir en se conformant aux lignes directrices suivantes.

- a) Les estimations qui figurent dans le corps d'un tableau statistique doivent être arrondies à la centaine près par la méthode d'arrondissement classique. Selon cette méthode, si le premier ou le seul chiffre à supprimer se situe entre 0 et 4, le dernier chiffre retenu ne change pas. Si le premier ou le seul chiffre à supprimer se situe entre 5 et 9, on augmente d'une unité (1) la valeur du dernier chiffre retenu. Par exemple, si l'on veut arrondir à la centaine près de la façon classique une estimation dont les deux derniers chiffres sont compris entre 00 et 49, il faut les remplacer par 00 et ne pas modifier le chiffre précédent (le chiffre des centaines). Si les deux derniers chiffres sont compris entre 50 et 99, il faut les remplacer par 00 et augmenter d'une unité (1) le chiffre précédent.
- b) Les totaux partiels de marge et les totaux de marge des tableaux statistiques doivent être calculés à partir de leurs éléments correspondants non arrondis, puis arrondis à leur tour à la centaine près selon la méthode d'arrondissement classique.
- c) Les moyennes, les proportions, les taux et les pourcentages doivent être calculés à partir d'éléments non arrondis (c'est-à-dire les numérateurs et (ou) dénominateurs), puis arrondis à une décimale par la méthode d'arrondissement classique. Si l'on veut arrondir une estimation à un seul chiffre décimal par cette méthode et que le dernier ou le seul chiffre à supprimer se situe entre 0 et 4, le dernier chiffre à retenir ne change pas. Si le premier ou le seul chiffre à supprimer se situe entre 5 et 9, on augmente d'une unité (1) le dernier chiffre à retenir.
- d) Les sommes et les différences d'agrégats (ou de rapports) doivent être calculées à partir de leurs éléments correspondants non arrondis, puis arrondies à leur tour à la centaine près (ou à la décimale près) selon la méthode d'arrondissement classique.
- e) Si, en raison de contraintes d'ordre technique ou autre, on applique une autre méthode que l'arrondissement classique, si bien que les estimations qui seront publiées ou diffusées de toute autre façon diffèrent des estimations correspondantes publiées par Statistique Canada et

- f) le NCHS, il est vivement conseillé à l'utilisateur d'indiquer la raison de ces divergences dans le ou les documents à publier ou à diffuser.

10.2 Lignes directrices pour la pondération de l'échantillon en vue de la totalisation

Le plan d'échantillonnage utilisé pour l'ECCES n'est pas autopondéré. Autrement dit, le poids d'échantillonnage n'est pas le même pour toutes les personnes qui font partie de l'échantillon. Même pour produire des estimations simples, y compris des tableaux statistiques ordinaires, l'utilisateur doit employer le poids d'échantillonnage approprié.

Sinon, les estimations calculées à partir des fichiers de microdonnées ne pourront être considérées comme représentatives de la population observée et ne correspondront pas à celles produites par Statistique Canada et par le NCHS.

L'utilisateur ne doit pas non plus perdre de vue qu'en raison du traitement accordé au poids, certains logiciels ne permettent pas d'obtenir des estimations qui coïncident exactement avec celles de Statistique Canada et du NCHS.

10.2.1 Définitions : estimations de type nominal, estimations quantitatives

Avant d'exposer la façon de totaliser et d'analyser les données de l'ECCES, il est bon de décrire les deux grandes catégories d'estimations ponctuelles des caractéristiques de la population qui peuvent être produites d'après le fichier de microdonnées de l'enquête.

Estimations de type nominal

Les estimations de type nominal sont des estimations du nombre ou du pourcentage de personnes qui, dans la population visée par l'enquête, possèdent certaines caractéristiques ou correspondent à une catégorie particulière. Le nombre de personnes qui fument tous les jours est un exemple d'estimation de ce genre. L'estimation du nombre de personnes qui possèdent une caractéristique particulière peut aussi être appelée « estimation d'un agrégat ».

Exemple de question de type nominal :

SMJ1_4 : **Actuellement, fumez-vous des cigarettes tous les jours, à l'occasion ou jamais?**

- | | |
|---|----------------|
| 1 | Tous les jours |
| 2 | À l'occasion |
| 3 | Jamais |

Estimations quantitatives

Les estimations quantitatives sont des estimations de totaux ou de moyennes, de médianes ou d'autres mesures de tendance centrale de quantités qui ont trait à tous les membres de la population observée ou à certains d'entre eux.

Un exemple d'estimation quantitative est le nombre moyen de cigarettes que fument par jour les personnes qui fument tous les jours. Le numérateur correspond à l'estimation du nombre total de cigarettes que fument par jour les personnes qui fument tous les jours et le dénominateur, à l'estimation du nombre de personnes qui fument tous les jours.

Exemple de question quantitative :

SMJ1_6 **Actuellement, combien de cigarettes fumez-vous chaque jour?**

|_|_| Nombre de cigarettes

10.2.2 Totalisation d'estimations de type nominal

On peut obtenir, à partir des fichiers de microdonnées, des estimations du nombre de personnes qui possèdent une caractéristique donnée en additionnant les poids finals de tous les enregistrements contenant des données sur la caractéristique étudiée.

Pour obtenir les proportions et les rapports de la forme \hat{X} / \hat{Y} , on doit :

- a) additionner les poids finals des enregistrements contenant la caractéristique voulue pour le numérateur (\hat{X});
- b) additionner les poids finals des enregistrements contenant la caractéristique voulue pour le dénominateur (\hat{Y});
- c) diviser l'estimation du numérateur par celle du dénominateur.

10.2.3 Totalisation d'estimations quantitatives

Pour obtenir des estimations quantitatives d'après le fichier de microdonnées, on doit :

- a) multiplier la valeur de la variable étudiée par le poids final, puis faire la somme de cette quantité pour tous les enregistrements visés pour obtenir le numérateur (\hat{X});

- b) faire la somme des poids finals des enregistrements contenant la variable étudiée pour obtenir le dénominateur (\hat{Y});
- c) diviser l'estimation du numérateur par l'estimation du dénominateur.

Par exemple, pour estimer le nombre moyen de cigarettes que fument chaque jour les personnes qui fument tous les jours, on multiplie la valeur déclarée pour la variable **SMJ1_6**¹³ par le poids, **WT_SAM**, puis on fait la somme des résultats pour tous les enregistrements pour lesquels la valeur de la variable **SMJ1_4** est « tous les jours » pour obtenir le numérateur (\hat{X}). Pour obtenir le dénominateur (\hat{Y}), on additionne les poids finals de tous les enregistrements pour lesquels la valeur de la variable **SMJ1_4** est « tous les jours ». Pour obtenir le nombre moyen de cigarettes fumées chaque jour par les personnes qui fument tous les jours, on divise (\hat{X}) par (\hat{Y}).

10.3 Lignes directrices pour l'analyse statistique

L'ECCES se fonde sur un plan de sondage complexe stratifié à plusieurs degrés et sur la sélection des répondants avec probabilités inégales. L'utilisation des données provenant d'une enquête aussi complexe pose des difficultés aux analystes, car le choix des méthodes d'estimation et de calcul de la variance dépend du plan de sondage et des probabilités de sélection.

Nombre de méthodes d'analyse intégrées aux progiciels statistiques permettent d'utiliser des poids, mais la signification et la définition de ces poids peuvent différer de celles applicables dans le contexte d'une enquête par sondage. Par conséquent, si les estimations calculées au moyen de ces progiciels sont souvent exactes, les variances n'ont, quant à elles, pratiquement aucune signification.

Dans le cas de nombreuses méthodes d'analyse (par exemple la régression linéaire, la régression logistique, l'analyse de la variance), on peut rendre l'application des progiciels courants plus significative en rééchantillonnant les poids qui figurent dans les enregistrements de façon à ce que le poids moyen soit égal à un (1). Les résultats produits par les progiciels classiques sont ainsi plus raisonnables puisque, même s'ils ne reflètent toujours pas la stratification et la mise en grappes du plan d'échantillonnage, ils tiennent compte de la sélection avec probabilités inégales. On peut effectuer cette transformation en divisant chaque poids utilisé dans l'analyse originale par la moyenne des poids originaux. Ainsi, la somme des poids originaux est égale à la taille de la population et la somme des poids rééchantillonnés est égale à la taille de l'échantillon.

10.4 Lignes directrices pour la diffusion

Avant de diffuser et(ou) de publier des estimations tirées du fichier de microdonnées à grande diffusion, l'utilisateur doit déterminer le nombre de personnes qui ont fourni les données entrant

¹³ Voir la section 11.2 pour les conventions appliquées pour nommer les variables.

dans le calcul de l'estimation. Si ce nombre est inférieur à 10, l'estimation pondérée ne doit pas être diffusée, quelle que soit la valeur de son coefficient de variation. Pour les estimations pondérées basées sur des échantillons d'au moins 10 personnes, l'utilisateur doit calculer le coefficient de variation de l'estimation et suivre les lignes directrices énoncées aux tableaux 10.1 et 10.2. Finalement, un avertissement : lorsqu'on rapporte des fréquences, il est important de noter qu'elles sont sous-estimées dû à la non-réponse et aux réponses inconnues.

Table 10.1 Lignes directrices concernant la variabilité d'échantillonnage suivies par Statistique Canada

| Type d'estimation | c.v. (en %) | Lignes directrices |
|-------------------|---------------------------|---|
| Acceptable | $0,0 \leq c.v. \leq 16,5$ | On peut envisager une diffusion générale non restreinte des estimations. Aucune annotation particulière n'est nécessaire. |
| Marginale | $16,5 < c.v. \leq 33,3$ | On peut envisager une diffusion générale non restreinte des estimations, en y joignant une mise en garde aux utilisateurs quant à la variabilité d'échantillonnage élevée liée aux estimations. Les estimations de ce genre doivent être identifiées par la lettre E (ou d'une autre manière similaire). |
| Inacceptable | $c.v. > 33,3$ | Statistique Canada recommande de ne pas publier des estimations dont la qualité est inacceptable. Toutefois, si l'utilisateur choisit de le faire, il doit alors adjoindre la lettre F (ou un autre identificateur semblable) et les diffuser avec l'avertissement suivant : « Nous avisons l'utilisateur que ...(précisez les données)... ne répondent pas aux normes de qualité de Statistique Canada pour ce programme statistique. Les conclusions tirées de ces données ne sauraient être fiables et seront fort probablement erronées. Ces données et toute conclusion qu'on pourrait en tirer ne doivent pas être publiées. Si l'utilisateur choisit de les publier, il est alors tenu de publier également le présent avertissement. » |

Tableau 10.2 Lignes directrices concernant la variabilité d'échantillonnage suivie par le NCHS

| Type d'estimation | c.v. (en %) | Lignes directrices |
|-------------------|------------------------------|--|
| Acceptable | $0 \leq \text{c.v.} \leq 30$ | On peut envisager une diffusion générale non restreinte des estimations. Aucune annotation particulière n'est nécessaire. |
| Marginale | $30 < \text{c.v.} \leq 50$ | Ces estimations peuvent être diffusées, mais il est recommandé d'ajouter une note en bas du tableau avertissant les utilisateurs que ce genre d'estimations ne répond pas aux normes de fiabilité du NCHS. |
| Inacceptable | $\text{c.v.} > 50$ | Le NCHS recommande de ne pas publier ce genre d'estimations. |

11. Utilisation du fichier

La présente section débute par l'examen de la **variable de pondération**. Vient ensuite une explication de la convention appliquée pour nommer les variables de l'ECCES. La section se termine par la description des diverses méthodes d'accès aux données offertes aux analystes.

11.1 Utilisation des poids

Le fichier ne contient qu'une seule variable de pondération, WT_SAM, qui est applicable aux répondants des deux pays. TOUTES LES VARIABLES DU FICHIER DEVRAIENT ÊTRE ANALYSÉES À L'AIDE DE CETTE VARIABLE DE PONDÉRATION.

(Pour une description plus détaillée du calcul de cette variable de pondération, voir la section 8 du présent guide.)

11.2 Convention appliquée pour nommer les variables

On a adopté une convention pour nommer les variables de l'ECCES qui permet aux utilisateurs des données de repérer et d'utiliser facilement celles-ci en fonction du module et du cycle. Les exigences qui suivent doivent être satisfaites : limiter les noms des variables à huit caractères au plus pour qu'il soit facile de les utiliser avec les logiciels d'analyse, préciser l'édition de l'enquête (cycle 1, 2 ...) dans le nom et permettre de repérer facilement les variables conceptuellement identiques dans l'éventualité d'un prochain cycle de l'enquête. Les noms des variables correspondant à des modules ou à des questions identiques ne devraient différer qu'en ce qui concerne la position réservée dans le nom à l'identification du cycle particulier durant lequel les données ont été recueillies.

11.2.1 Structure élémentaire des noms des variables de l'ECCES

Chacun des huit caractères du nom d'une variable fournit des renseignements sur le type de données que contient la variable.

| | |
|---------------------------|--------------------------------|
| Positions 1 et 2 : | Nom du module du questionnaire |
| Position 3 : | J: Identificateur de l'ECCES |
| Position 4 : | 1: Cycle de l'enquête |
| Position 5 : | Type de variable |
| Positions 6 à 8 : | Numéro de la question |

Par exemple, la structure du nom de la variable correspondant à la question 3D du module sur la limitation des activités, c'est-à-dire AHJ1_03D, est la suivante :

Positions 1 et 2 : **AH** Module de la limitation des activités
Positions 3 : **J** Identificateur de l'ECCES
Position 4 : **1** Cycle 1
Position 5 : **_** (_ = données recueillies)
Positions 6 à 8 : **03D** Numéro de la question et option de réponse

11.2.2 Positions 1 et 2 : Nom des modules du questionnaire

Les valeurs suivantes sont utilisées pour la composante du nom de la variable correspondant au module du questionnaire :

| | | | |
|----|---|----|---|
| AD | Administration | AH | Limitation des activités |
| DH | Variables démographiques et sur les ménages | PS | Test PAP |
| GH | État de santé général | MA | Mammographie |
| RA | Restriction des activités | DE | Visites chez le dentiste |
| CH | Problèmes de santé chroniques | IS | Assurance |
| DP | Dépression | RS | Limitation des activités professionnelles |
| CM | Utilisation des services - Santé mentale | SA | Satisfactions des patients |
| SM | Usage du tabac | PA | Activités physiques |
| HU | État de santé (Indice de l'état de santé - HUI) | SD | Renseignements socio-démographiques |
| HW | Taille et poids | IW | Revenu et patrimoine |
| HC | Utilisation des soins de santé | WT | Poids de l'échantillonnage |
| ME | Consommation de médicaments | | |

11.2.3 Position 3 : Identificateur de l'ECCES

La convention suivie pour identifier l'ECCES est semblable à celle utilisée pour les grandes enquêtes sur la santé de Statistique Canada, à savoir l'Enquête sur la santé dans les collectivités canadiennes et l'Enquête nationale sur la santé de la population (deux des trois enquêtes de base sur lesquelles s'appuie l'ECCES). Les noms de toutes les variables de l'ECCES contiennent un J en troisième position pour distinguer ces variables de celles correspondant aux questions provenant des enquêtes de base.

11.2.4 Position 4 : Cycle

Comme il s'agit du premier cycle de l'ECES, le nom de toutes les variables contient le chiffre 1 en quatrième position.

11.2.5 Position 5 : Type de variable

| _ | Variable collectée | Variable qui figure directement sur le questionnaire |
|----------|----------------------|---|
| D | Variable dérivée | Variable calculée d'après une ou plusieurs variables collectées ou codées, ordinairement pendant le traitement au Bureau central (p. ex., indice de l'état de santé) |
| F | Variable indicatrice | Variable calculée à partir d'une ou de plusieurs variables collectées (comme variable dérivée), mais ordinairement par l'application informatique de collecte des données, aux fins de son utilisation ultérieure durant l'interview (p. ex., indicateur de travail). |
| G | Variable groupée | Variables collectées, codées, supprimées ou dérivées agrégées en un groupe (p. ex., groupes d'âge) |

11.2.6 Positions 6 à 8 : Nom de la variable

En général, les trois dernières positions correspondent à la numérotation de la variable qui figure sur le questionnaire. On supprime la lettre « Q » utilisée pour représenter le mot « question » et on présente tous les numéros de question dans un format à deux chiffres. Par exemple, la question Q01A du questionnaire devient simplement 01A et la question Q15, simplement 15.

Dans le cas des questions pour lesquelles il existe plus d'une option de réponse, la position finale de la séquence dénommant la variable correspond à une lettre. Pour ce type de questions, on a créé de nouvelles variables pour faire la distinction entre une réponse « oui » ou « non » pour chaque option de réponse. Par exemple, s'il existe quatre options de réponse pour Q2, les nouvelles variables seront nommées Q2A pour l'option 1, Q2B pour l'option 2, Q2C pour l'option 3, etc. Si seules les options 2 et 3 ont été sélectionnées, alors Q2A = Non, Q2B = Oui, Q2C = Oui et Q2D = Non.

11.3 Accès aux données

Le fichier de microdonnées à grande diffusion de l'ECES est disponible gratuitement sur le site Web de Statistique Canada à http://www.statcan.ca/start_f.html et sur celui du National Center for Health Statistics of the U. S. Centers for Disease Control and Prevention à <http://www.cdc.gov/nchs/>.

Afin de respecter le droit à la vie privée des répondants qui participent à l'enquête, les fichiers de microdonnées doivent répondre à des normes rigoureuses de sécurité et de confidentialité, conformément à la *Loi sur la statistique* du Canada et à la *Public Health Service Act* des États-Unis avant de pouvoir faire l'objet d'une grande diffusion. Pour s'assurer du respect de ces normes, on soumet chaque fichier de microdonnées à un processus officiel d'examen destiné à confirmer qu'aucune personne ne pourra être identifiée. Les valeurs rares pour certaines variables susceptibles de permettre l'identification d'une personne sont supprimées du fichier ou agrégées en catégories moins détaillées, de façon à réduire au minimum le risque de divulgation de renseignements personnels. Parfois, ces variables sont les plus essentielles à l'analyse complète des données de l'enquête. Puisqu'une quantité importante de ressources est investie dans la collecte de ces données, il est important de prendre des mesures pour tirer le plein potentiel analytique des fichiers de microdonnées et bien rentabiliser l'investissement.

Statistique Canada et le NCHS ont mis en place des procédures visant à prévenir la divulgation de données confidentielles. L'information publiée ou autrement diffusée par les deux organismes fait l'objet d'un examen minutieux visant à assurer qu'aucune donnée confidentielle ne soit diffusée. Toute tentative en vue de déterminer l'identité d'un répondant individuel est strictement interdite. Tous les identificateurs directs, ainsi que toute caractéristique susceptible de faciliter la réidentification sont omis des fichiers de microdonnées à grande diffusion. Tout effort délibéré en vue de réidentifier les répondants auxquels correspondent les enregistrements contrevient aux garanties de confidentialité données aux fournisseurs de l'information. En utilisant les fichiers de données du Canada et des États-Unis, les utilisateurs acceptent de se conformer aux exigences suivantes :

1. les données des fichiers ne peuvent être utilisées qu'à des fins d'étude statistique et d'analyse des données;
2. aucune tentative ne doit être faite en vue d'identifier les personnes auxquelles correspondent les enregistrements figurant dans le fichier;
3. les fichiers de données ne doivent pas être couplés à d'autres données individuellement identifiables provenant d'autres sources canadiennes ou américaines.

Les analystes qui souhaitent utiliser des données recueillies aux États-Unis qui ont été supprimées pour des raisons de confidentialité peuvent avoir accès à certains fichiers de données non modifiés en s'adressant au Research Data Center (RDC) du NCHS. Ce centre, créé à l'intention des chercheurs venus de l'extérieur, est situé dans les locaux du NCHS, à Hyattsville, au Maryland. Les fichiers de données gardés au RDC peuvent aussi être consultés à distance par courriel. Les analystes qui souhaitent obtenir plus d'information sur la façon de faire une demande d'accès peuvent visiter le site Web suivant :

<http://www.cdc.gov/nchs/r&d/rdc.htm>.

Des frais sont perçus pour les services du RDC.