

**ENQUÊTE SUR LA SANTÉ DANS LES
COLLECTIVITÉS CANADIENNES (ESCC)
CYCLE 2.1 (2003)**

**GUIDE DU FICHIER DE MICRODONNÉES À
GRANDE DIFFUSION**

STATISTIQUE CANADA

JANVIER 2005

Table des matières

1.	INTRODUCTION	1
2.	CONTEXTE.....	2
3.	OBJECTIFS	4
4.	CONTENU DE L'ENQUÊTE.....	5
4.1	PROCESSUS DE CONSULTATION	5
4.2	CONTENU COMMUN	5
4.3	CONTENU OPTIONNEL.....	6
5.	PLAN D'ÉCHANTILLONAGE.....	8
5.1	POPULATION CIBLE.....	8
5.2	RÉGIONS SOCIO-SANITAIRES	8
5.3	TAILLE ET RÉPARTITION DE L'ÉCHANTILLON.....	9
5.4	BASES DE SONDAGE ET STRATÉGIES D'ÉCHANTILLONNAGE DES MÉNAGES	9
5.4.1	ÉCHANTILLONNAGE DES MÉNAGES À PARTIR DE LA BASE ARÉOLAIRE	9
5.4.2	ÉCHANTILLONNAGE DES MÉNAGES À PARTIR DE LA BASE LISTE DE NUMÉROS DE TÉLÉPHONE	12
5.4.3	ÉCHANTILLONNAGE DES MÉNAGES À PARTIR DE LA BASE DE SONDAGE À CA DE NUMÉROS DE TÉLÉPHONE	12
5.5	ÉCHANTILLONNAGE DES PERSONNES INTERVIEWÉES	13
5.6	RÉPARTITION DE L'ÉCHANTILLON SUR LA PÉRIODE DE COLLECTE DES DONNÉES	14
5.7	ACHAT D'UNITÉS D'ÉCHANTILLONNAGE SUPPLÉMENTAIRES DANS TROIS RSS DE LA PROVINCE DE QUÉBEC	14
5.8	ÉTUDE SPÉCIALE SUR LE MODE DE COLLECTE (IPAO VERSUS ITAO).....	15
6.	COLLECTE DES DONNÉES	18
6.1	DÉVELOPPEMENT DU QUESTIONNAIRE ET MÉTHODE DE COLLECTE DES DONNÉES	18
6.2	SUPERVISION ET CONTRÔLE	18
6.3	ESSAIS SUR LE TERRAIN.....	19
6.4	TECHNIQUES D'INTERVIEW	19
6.5	RÉDUCTION DE LA NON-RÉPONSE	19
6.6	RÉSPECT DE LA VIE PRIVÉE.....	20
7.	TRAITEMENT DES DONNÉES	21
7.1	VÉRIFICATION	21
7.2	CODAGE	21
7.3	CRÉATION DE VARIABLES DÉRIVÉES ET GROUPEES	21
7.4	PONDÉRATION	22
7.5	ÉLIMINATION DES RENSEIGNEMENTS CONFIDENTIELS	22
8.	PONDÉRATION	23
8.1	PONDÉRATION DE L'ÉCHANTILLON	23
8.1.1	PONDÉRATION DE L'ÉCHANTILLON PROVENANT DE LA BASE ARÉOLAIRE.....	24
8.1.2	PONDÉRATION DE L'ÉCHANTILLON PROVENANT DE LA BASE TÉLÉPHONIQUE.....	27
8.1.3	INTÉGRATION DES BASES ARÉOLAIRE ET TÉLÉPHONIQUE (I1).....	31
8.1.4	EFFET SAISONNIER (I2).....	31
8.1.5	POSTSTRATIFICATION (I3)	32
8.1.6	PARTICULARITÉS DE LA PONDÉRATION POUR LES TROIS TERRITOIRES.....	32
8.1.7	PARTICULARITÉS DE LA PONDÉRATION POUR LES RSS OÙ A EU LIEU L'ÉTUDE DU MODE DE COLLECTE	33

9.	QUALITÉ DES DONNÉES	36
9.1	TAUX DE RÉPONSE	36
9.2	ERREURS DANS LES ENQUÊTES	42
9.2.1	ERREURS NON DUES À L'ÉCHANTILLONNAGE	42
9.2.2	ERREURS DUES À L'ÉCHANTILLONNAGE	42
10.	LIGNES DIRECTRICES POUR LA TOTALISATION, L'ANALYSE ET LA DIFFUSION	44
10.1	LIGNES DIRECTRICES POUR L'ARRONDISSEMENT	44
10.2	LIGNES DIRECTRICES POUR LA PONDÉRATION DE L'ÉCHANTILLON EN VUE DE LA TOTALISATION	45
10.2.1	DÉFINITIONS DES CATÉGORIES D'ESTIMATIONS : DE TYPE NOMINAL PAR OPPOSITION À QUANTITATIVES	45
10.2.2	TOTALISATION D'ESTIMATIONS DE TYPE NOMINAL	46
10.2.3	TOTALISATION D'ESTIMATIONS QUANTITATIVES	46
10.3	LIGNES DIRECTRICES POUR L'ANALYSE STATISTIQUE	47
10.4	LIGNES DIRECTRICES POUR LA DIFFUSION	47
11.	TABLEAUX DE LA VARIABILITÉ D'ÉCHANTILLONNAGE APPROXIMATIVE	49
11.1	COMMENT UTILISER LES TABLEAUX DE CV POUR LES ESTIMATIONS DE TYPE NOMINAL	49
11.2	EXEMPLES D'UTILISATION DES TABLEAUX DE CV POUR DES ESTIMATIONS DE TYPE NOMINAL	51
11.3	COMMENT UTILISER LES TABLEAUX DE CV POUR CALCULER LES LIMITES DE CONFIANCE	54
11.4	EXEMPLE D'UTILISATION DE TABLEAUX DE CV POUR OBTENIR DES LIMITES DE CONFIANCE	55
11.5	COMMENT UTILISER LES TABLEAUX DE CV POUR EFFECTUER UN TEST Z	56
11.6	EXEMPLE D'UTILISATION DES TABLEAUX DE CV POUR EFFECTUER UN TEST Z	56
11.7	VARIANCES OU COEFFICIENTS DE VARIATION EXACTS	56
11.8	SEUILS POUR LA DIFFUSION DES ESTIMATIONS RELATIVES À L'ESCC	58
12.	UTILISATION DU FICHIER	59
12.1	UTILISATION DE LA VARIABLE DE PONDÉRATION	59
12.2	CONVENTION APPLIQUÉE POUR NOMMER LES VARIABLES	59
12.2.1	STRUCTURE ÉLÉMENTAIRE DES NOMS DES VARIABLES DE L'ESCC	59
12.2.2	POSITIONS 1 À 3 : NOM DE LA VARIABLE/SECTION DU QUESTIONNAIRE	60
12.2.3	POSITION 4 : CYCLE	61
12.2.4	POSITION 5 : TYPE DE VARIABLE	62
12.2.5	POSITIONS 6 À 8 : NOM DE LA VARIABLE	62
12.3	ACCÈS AU FICHIER MAÎTRE	62

Liste des annexes

Annexe A : [Questionnaire](#)

Annexe B : [Cliché d'enregistrement](#)

Annexe C : [Dictionnaire des données](#)

Annexe D : [Variables dérivées et groupées](#)

Annexe E : [Tableaux de CV](#)

1. Introduction

L'Enquête sur la santé dans les collectivités canadiennes (l'ESCC) est une enquête transversale qui vise à recueillir des renseignements sur l'état de santé, l'utilisation des services de santé et les déterminants de la santé de la population canadienne. Le cycle de collecte des données de l'ESCC s'étend sur deux années. La première année du cycle, indiquée par la notation « .1 », correspond à une enquête générale sur la santé de la population réalisée auprès d'un grand échantillon et conçue pour fournir des estimations fiables à l'échelle de la région sociosanitaire. La deuxième année du cycle, représentée par la notation « .2 », correspond à une enquête de moins grande portée conçue pour fournir des données à l'échelle provinciale sur des sujets particuliers ayant trait à la santé.

Le présent fichier de microdonnées contient les données du troisième cycle de l'ESCC (cycle 2.1). Les renseignements ont été recueillis de janvier 2003 à décembre 2003 pour 126 régions sociosanitaires couvrant les dix provinces et les trois territoires. Les données du cycle 2.1 de l'ESCC sont recueillies auprès des personnes de 12 ans et plus vivant dans des logements privés. Sont exclues de la base de sondage les personnes vivant sur les réserves indiennes et les terres de la Couronne, les résidents des établissements, les membres à temps plein des Forces canadiennes et les personnes vivant dans certaines régions éloignées. L'ESCC couvre environ 98 % de la population canadienne âgée de 12 ans et plus.

Le présent document a pour but de faciliter la manipulation du fichier de microdonnées du cycle 2.1 de l'ESCC qui est décrit en détails dans le texte et les annexes qui suivent.

Pour toute question concernant les ensembles de données ou leur utilisation, s'adresser à :

Service d'aide aux utilisateurs des produits électroniques : 1 (800) 949-9491

Totalisations spéciales ou renseignements généraux sur les données :

Services personnalisés à la clientèle, Division de la statistique de la santé : (613) 951-1746

Courriel : hd-ds@statcan.ca

Renseignements sur le télé-accès :

(613) 951-1653

Courriel :

cchs-escc@statcan.ca

Télécopieur :

(613) 951-4198

2. Contexte

En 1991, le Groupe de travail national sur l'information en matière de santé a relevé plusieurs problèmes posés par le système d'information sur la santé. Selon ses membres, les données étaient fragmentées, elles étaient incomplètes, elles ne pouvaient être partagées facilement et elles n'étaient pas analysées aussi pleinement que possible; en outre, les résultats des études réalisées n'atteignaient pas de façon régulière la population canadienne¹. Pour résoudre ces problèmes, l'Institut canadien d'information sur la santé (ICIS), Statistique Canada et Santé Canada ont conjugué leurs efforts en vue de créer un Carnet de route de l'information sur la santé.

L'Initiative du Carnet de route a été lancée en réponse directe aux préoccupations et aux souhaits exprimés par plus de 500 personnes représentant un large éventail d'organismes et de groupes d'intérêt. Au début de 1999, les trois organismes nationaux susmentionnés ont mené une consultation nationale à grande échelle sur les besoins d'information en matière de santé. Les participants ont insisté sur le fait que les organismes nationaux doivent collaborer en vue de renforcer le système canadien d'information sur la santé et mettre à profit les investissements et les compétences considérables aux niveaux local, régional et provincial/territorial².

Le Carnet de route représente une contribution importante à l'édification d'un système national complet d'information sur la santé et de l'infrastructure requise pour donner aux Canadiens l'information dont ils ont besoin pour entretenir et améliorer le système de santé et la santé de la population du Canada³. Un plan d'action coordonné est requis. Le gouvernement seul ou une seule organisation ne peut pas lutter contre les problèmes mentionnés plus haut. La collaboration à tous les niveaux — organismes de santé nationaux, provinciaux, territoriaux, régionaux et locaux — est une condition préalable au succès⁴.

Notre système d'information sur la santé devrait nous fournir l'information pour répondre aux questions cruciales ci-dessous⁵ :

1. À quel point le système de santé est-il sain?
2. À quel point les Canadiens sont-ils en santé?

La première question englobe l'efficacité, l'efficience et la réceptivité du système de santé. En règle générale, un système de santé efficace, efficient et réceptif est un système qui offre aux Canadiens les soins de qualité auxquels ils s'attendent⁶.

La deuxième question est plus générale et traite des objectifs de base du système : la santé des Canadiens s'améliore-t-elle? Afin de répondre à cette question et à d'autres aussi importantes,

¹ 1999. Carnet de route de l'information sur la santé — Répondre aux besoins, Santé Canada, Statistique Canada. p. 3.

² 1999. Ibid. p. 1.

³ 1999. Ibid. p. 1.

⁴ 1999. Ibid. p. 3.

⁵ 1999. Ibid. p. 3.

⁶ 1999. Ibid. p. 3.

nous avons besoin d'un système solide d'information sur la santé⁷. Ce système doit posséder six grandes caractéristiques⁸. Il doit être :

- sécuritaire et respecter le droit des Canadiens à la vie privée,
- cohérent,
- pertinent,
- intégrable,
- flexible,
- convivial et accessible.

Ce nouveau système d'information sur la santé doit être à jour, fournir des renseignements orientés vers la personne et s'appuyer sur des normes de données communes à d'autres enquêtes sur la santé de la population canadienne, telles que l'Enquête nationale sur la santé de la population (ENSP). Il doit également fournir de nouveaux ensembles de données ou des ensembles de données étoffées, des données sur les services de santé, des données sur les résultats relatifs à la santé, l'état de santé et les déterminants non médicaux de la santé, des données sur les résultats d'interventions particulières, des études spéciales portant sur des questions prioritaires, des données sur les coûts selon le service, des protocoles d'échange de données, une plus grande capacité d'analyse des données, ainsi que des rapports publics sur le système de santé⁹.

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) a été conçue compte tenu de ce mandat. Le format, le contenu et les objectifs de cette enquête ont été définis après avoir mené des consultations approfondies auprès de spécialistes et d'intervenants fédéraux, provinciaux et communautaires en vue de déterminer leurs exigences en matière de données¹⁰.

Le présent Guide du fichier de microdonnées à grande diffusion est publié en réponse à l'exigence de recueillir des données fiables et pertinentes sur les services de santé, l'état de santé et les questions relatives à la santé revêtant une importance pour la population canadienne — à l'échelle régionale, provinciale et nationale — et de diffuser cette information au public.

⁷ 1999. Ibid. p. 5.

⁸ Ces caractéristiques sont décrites en détail dans le document intitulé Carnet de route de l'information sur la santé : Répondre aux besoins, 1999, Institut canadien d'information sur la santé. ISBN 1-895581-30-3. (<http://www.cihi.ca>)

⁹ 1999. Ibid. p. 11-14.

¹⁰ 1999. Initiative du carnet de route ... Lancer le processus. Institut canadien d'information sur la santé/Statistique Canada. ISBN 1-895581-70-2. p. 19.

3. Objectifs

Les objectifs principaux de l'ESCC sont les suivants :

- fournir des estimations transversales à jour et fiables des déterminants de la santé, de l'état de santé et de l'utilisation des services de santé à travers le Canada,
- recueillir des données à l'échelle infraprovinciale,
- créer un instrument d'enquête souple permettant :
 - de combler des lacunes statistiques particulières à l'échelle de la région sociosanitaire,
 - d'élaborer un contenu d'enquête thématique en vue de recueillir des données importantes, et
 - de répondre aux nouvelles questions ayant trait à la santé et aux services de santé à mesure qu'elles surviennent.

L'ESCC, en tant que composante importante du Programme des enquêtes sur la santé de Statistique Canada, permet de combler des besoins d'information accrus en matière de santé. Il s'agit de :

- faciliter l'élaboration de politiques gouvernementales,
- fournir des données permettant de réaliser des études analytiques qui aideront à comprendre les déterminants de la santé,
- recueillir des données sur les corrélations entre la santé et les facteurs économiques, sociaux, démographiques, professionnels et environnementaux,
- permettre de mieux comprendre la relation entre l'état de santé et l'utilisation des services de santé.

4. Contenu de l'enquête

La présente section décrit le processus général de consultation suivi pour élaborer le contenu de l'enquête et résume le contenu choisi pour l'étude. La deuxième sous-section décrit en détail le contenu commun, suivie d'une sous-section expliquant le contenu optionnel de l'ESCC (cycle 2.1).

4.1 Processus de consultation

L'un des objectifs principaux de l'ESCC est de combler les lacunes statistiques en ce qui concerne les déterminants de la santé, l'état de santé et l'utilisation des services de santé, à l'échelle de la région sociosanitaire.

Pour identifier ces lacunes, des consultations ont été menées à l'automne 2001 auprès de plus de 200 représentants d'organismes gouvernementaux régionaux, provinciaux et fédéraux, ainsi qu'auprès de chercheurs spécialisés dans l'étude de la santé de la population.

Alors que les consultations menées avant le cycle 1.1 de l'ESCC s'appuyaient sur une combinaison de méthodes qualitatives et quantitatives en vue de déterminer la priorité relative de grands domaines thématiques, l'objectif principal des consultations concernant le cycle 2.1 était de cerner les nouveaux domaines pour lesquels existent des lacunes statistiques.

À la suite de ces consultations, une liste de sujets à inclure dans le cycle 2.1 de l'enquête a été dressée par Statistique Canada et approuvée par un comité consultatif formé de représentants des régions sociosanitaires, des ministères provinciaux et territoriaux de la Santé et de Santé Canada.

Le questionnaire final du cycle 2.1 de l'ESCC comprend les éléments suivants : environ 25 minutes de contenu commun, s'adressant à tous les répondants, environ 5 minutes de contenu pour les sous-échantillons, c'est-à-dire certains modules du questionnaire que l'on a posé uniquement à un nombre de répondants suffisant pour produire des estimations fiables aux échelles nationale et provinciale, et environ 10 minutes de contenu optionnel.

Chaque région sociosanitaire a eu droit à dix minutes de contenu optionnel. Les représentants régionaux ont choisi les modules du questionnaire d'après une liste pré-établie, en se fondant sur les besoins et les priorités à l'échelle locale. Les modules de contenu optionnel n'ont été posés qu'aux répondants vivant dans les régions sociosanitaires les ayant sélectionnés.

4.2 Contenu commun

Les sujets formant le contenu commun sont très variés: ils vont de la consommation d'alcool à l'exposition à la fumée secondaire, en passant par l'activité physique et l'incapacité au cours des deux dernières semaines. Le tableau 4.1 présente le contenu commun du cycle 2.1 de l'ESCC tel qu'il a été déterminé lors des consultations. Les questions de l'enquête portant sur les sujets du contenu commun ont été posées à tous les répondants dans toutes les régions sociosanitaires.

Tableau 4.1 Cycle 2.1 de l'ESCC – Modules de contenu commun

Consommation d'alcool Problèmes de santé chroniques Changements pour améliorer la santé Exposition à la fumée des autres Insécurité alimentaire Vaccination contre la grippe Consommation de fruits et de légumes État de santé général Utilisation des soins de santé Taille et poids Soins à domicile Blessures Mammographie Expériences maternelles Santé bucco-dentaire 1 Test Papanicolaou Activités physiques	Mouvement répétitif Limitation des activités Comportement sexuel Usage du tabac Incapacité au cours des deux dernières semaines Organismes bénévoles Usage du tabac chez les jeunes Niveau de scolarité Identificateurs géographiques Données démographiques et composition du ménage Revenu Couverture par une assurance Population active Renseignements sociodémographiques
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4.3 Contenu optionnel

Le contenu thématique des modules optionnels a également été déterminé durant le processus de consultation (voir le tableau 4.2). Cependant, les sujets qui entrent dans cette catégorie ont été considérés comme optionnels de sorte que les régions ayant besoin de données ou s'intéressant à certains sujets, puissent choisir les modules de contenu pertinents à intégrer dans le questionnaire du cycle 2.1 de l'ESCC leur étant destiné. L'avantage de cette approche est que les régions sociosanitaires peuvent adapter la couverture des sujets ayant trait à la santé en fonction de leurs caractéristiques. L'inconvénient est que, contrairement à celles des modules formant le volet du contenu commun, les données provenant des modules de contenu optionnel ne peuvent être généralisées facilement à l'échelle du Canada. Par conséquent, la taille et les caractéristiques des régions dans lesquelles les modules optionnels sont utilisés limitent les comparaisons interrégionales des résultats.

Tableau 4.2 Cycle 2.1 de l'ESCC – Modules de contenu optionnel

Dépendance à l'égard de l'alcool Tension artérielle Examen des seins Auto-examen des seins Dépistage du cancer du côlon et du rectum Consultations des spécialistes de la santé mentale Visites chez le dentiste * Dépression Utilisation de compléments vitaminiques Détresse Conduite automobile et sécurité * Troubles de l'alimentation Choix alimentaires Satisfaction à l'égard du système de soins de santé État de santé - SF-36 Indice de l'état de santé (HUI) * Sécurité à la maison Drogues illicites Activités de loisir Contrôle	Consommation de médicaments * Dépendance à la nicotine Santé bucco-dentaire 2 * Satisfaction des patients Examen général Consultation d'un médecin (usage du tabac) Jeu pathologique Test de l'antigène spécifique prostatique Satisfaction concernant la disponibilité des services de santé Satisfaction à l'égard de la vie Activités sédentaires Estime de soi Outils pour arrêter de fumer Soutien social Étapes du changement (usage du tabac) Pensées suicidaires et tentatives de suicide Variantes du tabagisme Utilisation de protections Stress au travail
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

* Prévu pour le contenu pour les sous-échantillons, mais pouvait aussi être choisi par les régions sociosanitaires à titre de contenu optionnel.

5. Plan d'échantillonnage

5.1 Population cible

L'ESCC vise la population de 12 ans et plus vivant à domicile des dix provinces et des trois territoires. Sont exclues du champ de l'enquête les personnes vivant sur les réserves indiennes et les terres de la Couronne, les résidents des établissements, les membres à temps plein des Forces canadiennes et les personnes vivant dans certaines régions éloignées. L'ESCC couvre environ 98 % de la population canadienne de 12 ans et plus.

5.2 Régions sociosanitaires

À des fins administratives, chaque province est divisée en plusieurs régions sociosanitaires (RSS) et chaque territoire est considéré comme formant une RSS unique (tableau 5.1). En collaboration avec les provinces, Statistique Canada a modifié légèrement les limites de certaines RSS afin qu'elles correspondent aux données géographiques du Recensement de 2001. Durant le cycle 2.1 de l'ESCC, des données ont été recueillies pour 123 RSS dans les 10 provinces, ainsi que pour une RSS par territoire, soit, en tout, pour 126 RSS.

Tableau 5.1. Nombre de régions sociosanitaires et tailles visées d'échantillon selon la province/territoire

Province	Nombre de RSS	Taille totale de l'échantillon (visée)
Terre-Neuve et Labrador	6	4 010
Île-du-Prince-Édouard	4	2 000
Nouvelle-Écosse	6	5 040
Nouveau-Brunswick	7	5 150
Québec	17	24 280
Ontario	37	42 260
Manitoba	10	7 500
Saskatchewan	11	7 720
Alberta	9	14 200
Colombie-Britannique	16	16 090
Yukon	1	850
Territoires du Nord-Ouest	1	900
Nunavut	1	700
Canada	126	130 700

5.3 Taille et répartition de l'échantillon

Afin de produire des estimations fiables pour les 126 RSS et compte tenu du budget accordé pour le cycle 2.1 de l'ESCC, il a été établi que cette composante de l'enquête devrait être réalisée auprès d'un échantillon de 130 700 personnes. La production d'estimations fiables pour la RSS était l'objectif primordial, mais la qualité des estimations pour certaines caractéristiques importantes pour les provinces a été jugée importante également. Par conséquent, la stratégie de répartition de l'échantillon, qui comporte trois étapes, accorde une importance plus ou moins égale aux RSS et aux provinces. Lors des deux premières étapes, l'échantillon a été réparti entre les provinces en fonction de leur compte de population et du nombre de RSS qu'elles contiennent (tableau 5.1). À la troisième étape, chaque échantillon provincial a été réparti entre les RSS proportionnellement à la racine carrée de la population estimée de la RSS.

Cette stratégie en trois étapes permet d'obtenir un échantillon suffisant pour chaque RSS, sans perturber considérablement la répartition interprovinciale de l'échantillon. L'effectif des échantillons a été augmenté avant la collecte des données afin de tenir compte des logements hors du champ de l'enquête ou vacants, et du taux prévu de non-réponse. Pour la liste complète des RSS et des tailles finales d'échantillons, consulter la section 9 sur la qualité des données.

Il convient de souligner que les trois territoires, qui ont été traités séparément, n'étaient pas visés par la stratégie susmentionnée de répartition de l'échantillon. En tout, 850 unités d'échantillonnage ont été attribuées au Yukon, 900 aux Territoires du Nord-Ouest et 700 au Nunavut.

5.4 Bases de sondage et stratégies d'échantillonnage des ménages

L'échantillon de ménages du cycle 2.1 de l'ESCC a été sélectionné à partir de trois bases de sondage : 48 % de l'échantillon de ménages provenait d'une base de sondage aréolaire, 50 % provenait d'une base liste de numéros de téléphone et 2 % provenait d'une base de sondage à composition aléatoire (CA).

5.4.1 Échantillonnage des ménages à partir de la base aréolaire

La base aréolaire conçue pour l'Enquête sur la population active (EPA) du Canada a servi de base de sondage principale pour l'ESCC. Le plan d'échantillonnage de l'EPA est un plan d'échantillonnage en grappes stratifié à plusieurs degrés où le logement représente l'unité finale d'échantillonnage¹¹. À la première étape, on a formé des strates homogènes et sélectionné un échantillon indépendant de grappes, dans chaque strate. À la deuxième étape, on a dressé la liste des logements pour chaque grappe, puis on a sélectionné les logements, ou les ménages, d'après les listes.

Pour les besoins du plan d'échantillonnage, chaque province est répartie en trois catégories de région, à savoir les grands centres urbains, les villes et les régions rurales. Des strates

¹¹ Statistique Canada (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Statistique Canada, numéro 71-526-XPB au catalogue.

géographiques ou socioéconomiques sont formées à l'intérieur de chaque grand centre urbain. Dans les strates, des grappes sont formées par regroupement de 150 à 250 logements. Dans certains centres urbains, des strates distinctes sont créées pour les immeubles à appartements ou les secteurs de dénombrement (SD) pour lesquels le revenu moyen du ménage est élevé. Dans chaque strate, on sélectionne six grappes ou immeubles résidentiels (pouvant compter de 12 à 18 appartements) par une méthode d'échantillonnage aléatoire avec probabilité proportionnelle à la taille (PPT), cette dernière correspondant au nombre de ménages. Le nombre 6 est utilisé pour l'ensemble du plan d'échantillonnage afin de permettre le renouvellement mensuel d'un sixième de l'échantillon de l'EPA.

Les autres villes et régions rurales de chaque province sont stratifiées, en premier lieu, en fonction de données géographiques, puis selon les caractéristiques socioéconomiques. Dans la plupart des strates, on sélectionne six grappes (habituellement des SD) par la méthode PPT. Pour les strates où la densité de population est faible, on suit un plan en trois étapes en vertu duquel on sélectionne deux ou trois unités primaires d'échantillonnage (UPE), qui correspondent normalement à des groupes de SD, puis on les répartit en grappes dont six sont sélectionnées pour faire partie de l'échantillon. La sélection est réalisée à chaque étape selon la méthode PPT.

Une fois que la liste des nouvelles grappes est établie, on obtient l'échantillon par échantillonnage systématique des logements. Le tableau 5.2 donne un aperçu des catégories d'UPE utilisées pour l'échantillon complet de l'EPA. Le *rendement* correspond au nombre de ménages sélectionnés dans le cadre de l'EPA pour un mois donné. Comme les taux d'échantillonnage sont prédéterminés, il existe souvent un écart entre la taille prévue d'échantillon et les chiffres obtenus. Ainsi, le rendement de l'échantillon est parfois excessif. Cette situation se présente surtout pour les secteurs où le nombre de logements a augmenté à la suite de nouveaux projets de construction, par exemple. Pour réduire le coût de la collecte des données, on corrige la production excessive par élimination, dès le départ, d'une partie des unités sélectionnées et modifications du coefficient de pondération appliqué dans le plan d'échantillonnage. Cette opération, habituellement réalisée au niveau agrégé, porte le nom de *stabilisation de l'échantillon*. En plus, on augmente la taille requise de l'échantillon de ménages pour tenir compte du fait qu'en général, environ 12 % de logements ne sont pas occupés par des ménages faisant partie du champ d'observation (certains logements sont vacants ou occupés de façon saisonnière, et d'autres sont occupés par des ménages non visés par l'enquête).

Tableau 5.2 Unité primaire d'échantillonnage, taille et rendement

Région	Unité primaire d'échantillonnage (UPE)	Taille (ménages par UPE)	Rendement (ménages échantillonnés)
Toronto, Montréal, Vancouver	Grappe	200 à 250	6
Autres villes	Grappe	150 à 200	8
Base des immeubles à appartements	Appartement	Varie	5
La plupart des régions rurales/petits centres urbains	Secteur de dénombrement	300	10

Afin de répondre aux exigences particulières à l'ESCC, certaines modifications ont dû être apportées à cette stratégie d'échantillonnage¹². Pour obtenir un échantillon de base de 62 000 ménages pour l'ESCC, il faut sélectionner 84 000 logements d'après la base aréolaire (pour tenir compte des logements vacants et des ménages non-répondants). Chaque mois, le plan d'échantillonnage de l'EPA fournit environ 68 000 logements répartis entre les diverses régions économiques du Canada, alors que, pour l'ESCC, il faut obtenir un échantillon total de 84 000 logements répartis entre les RSS, dont les limites géographiques diffèrent de celles des régions économiques de l'EPA. Globalement, l'ESCC nécessite environ 24 % plus de logements que le nombre produit par le mécanisme de sélection de l'EPA, ce qui correspond à un facteur de redressement de 1,24 (84 000/68 000). Toutefois, comme ce facteur de redressement varie de 0,6 à 6,0 au niveau de la RSS, certains ajustements sont nécessaires.

Les modifications apportées au processus de sélection dans une RSS varient selon la grandeur du facteur de redressement. Pour les RSS pour lesquelles le facteur est égal ou inférieur à 1, on procède à une simple stabilisation de l'échantillon de logements, telle que décrite plus haut. Pour celles pour lesquelles le facteur est supérieur à 1, mais inférieur ou égal à 2, on répète le processus d'échantillonnage des logements à l'intérieur d'une UPE pour toutes les UPE sélectionnées appartenant à la RSS en question. Pour les RSS pour lesquelles le facteur est supérieur à 2, mais inférieur ou égal à 4, on répète le processus d'échantillonnage des UPE ainsi que celui des logements dans les UPE. Pour les RSS pour lesquelles le facteur est compris entre 4 et 6, on répète le processus d'échantillonnage des UPE deux fois au lieu d'une, et celui de sélection des logements une fois uniquement. Dans les cas où la méthode choisie produit un excédent inutile de logements, on procède à la stabilisation de l'échantillon.

Il convient de souligner que les modifications apportées au processus d'échantillonnage de l'EPA aboutissent, au plus, au triplement du nombre d'UPE sélectionnées et, au plus, au doublement du nombre de logements sélectionnés dans les UPE, ce qui explique la valeur maximale de 6,0 du facteur de redressement. Pour les RSS, on a plafonné délibérément la valeur du facteur de redressement à 6,0 pour deux raisons : pour limiter le nombre de listes de grappes à produire (chaque nouvelle UPE sélectionnée nécessite une liste) et pour éviter les effets éventuels de grappes dus à la sélection d'un trop grand nombre de logements dans une même UPE. Cette limite du facteur de redressement appliqué pour certaines RSS a conséquemment dicté le nombre de ménages qu'il fallait sélectionner à partir des bases de sondage téléphoniques.

Échantillonnage des ménages à partir de la base aréolaire dans les trois territoires

Pour des raisons opérationnelles, le plan d'échantillonnage à partir de la base aréolaire utilisé pour les trois territoires comporte une étape supplémentaire. Pour chaque territoire, on a commencé par stratifier les collectivités (municipalités) faisant partie du champ de l'enquête en fonction de diverses caractéristiques (population, données géographiques, proportion d'Inuit et/ou d'Autochtones et revenu médian du ménage). On a défini de la sorte 5 strates pour le Yukon, 10 pour les Territoires du Nord-Ouest et 6 pour le Nunavut. Puis, le premier degré d'échantillonnage a consisté à sélectionner aléatoirement une collectivité avec probabilité proportionnelle à la taille

¹² Morano M., Lessard, S. et Béland, Y. (2000). Creation of a dual frame for the Canadian Community Health Survey, 2000 *Proceedings of the Survey Methods Section*, Ottawa: La Société statistique du Canada, 249-254.

de la population dans chaque strate définie. Puis, de là, on a appliqué, à l'intérieur de chaque collectivité, une stratégie d'échantillonnage des ménages à partir de la base aréolaire identique à celle décrite plus haut.

Il convient de mentionner que la base de sondage de l'ESCC couvre 90 % des ménages privés du Yukon, 97 % de ceux des Territoires du Nord-Ouest et 71 % de ceux du Nunavut.

5.4.2 Échantillonnage des ménages à partir de la base liste de numéros de téléphone

À l'exception de 8 RSS, la base liste de numéros de téléphone a été utilisée dans toutes les régions; seules les 5 RSS qui ont eu recours à la base CA et les trois territoires n'ont pas utilisé la base liste. À l'instar de la base de sondage à CA, on a utilisé une base liste de numéros de téléphone pour compléter la base aréolaire. À cette fin, on a couplé l'annuaire téléphonique du Canada, un disque compact disponible sur le marché contenant les noms, les adresses et les numéros de téléphone répertoriés dans les annuaires de téléphone du Canada, à des fichiers internes administratifs de conversion afin d'obtenir les codes postaux que l'on a fait correspondre aux RSS afin de créer des listes stratifiées de numéros de téléphone, à raison d'une liste par RSS. Dans chaque strate, on a sélectionné le nombre requis de numéros de téléphone d'après la base liste par échantillonnage aléatoire simple. Comme pour la base de sondage à CA, on a sélectionné des numéros de téléphone supplémentaires pour tenir compte des numéros hors service ou hors du champ d'observation. Le taux de réussite enregistré pour la sélection d'après la liste de numéros de téléphone est nettement plus élevé que celui observé pour la liste établie par CA, variant de 70 % à 80 %.

Il importe de souligner que la couverture de la base liste de numéros de téléphone est moins importante que celle de la base de sondage à CA, car les numéros non publiés n'ont aucune chance d'être sélectionnés. Néanmoins, comme la liste des numéros de téléphone n'a été utilisée que pour les RSS pour lesquelles la base aréolaire était la source principale de l'échantillon, l'effet du sous-dénombrement dû à l'utilisation de la base liste de numéros de téléphone a été minimal et a pu être corrigé par pondération.

5.4.3 Échantillonnage des ménages à partir de la base de sondage à CA de numéros de téléphone

Dans certaines RSS, on a utilisé pour certains mois de collecte, un échantillon de numéros de téléphone provenant de la base de sondage à composition aléatoire (CA) en plus de la base aréolaire. L'échantillonnage de ménages à partir de la base à CA a été réalisé selon la méthode d'élimination des banques non valides (EBNV) adoptée par l'Enquête sociale générale¹³. Une banque de cent numéros (c'est-à-dire les huit premiers chiffres d'un numéro de téléphone à 10 chiffres) est considérée comme non valide si elle ne contient aucun numéro de téléphone résidentiel. Au départ, la base de sondage comprend la liste de toutes les banques valides de cent numéros et celles qui ne sont pas valides sont éliminées de la base de sondage à mesure qu'on les

¹³ Norris, D.A., Paton, D.G. (1991), L'Enquête sociale générale canadienne: bilan des cinq premières années. *Techniques d'enquête* (Statistique Canada, Catalogue 12-001); 17, pp. 245-260.

repère. Il convient de souligner que ces banques de cent numéros ne sont éliminées de la base de sondage que lorsque l'on possède des preuves qu'elles ne sont pas valides provenant de sources diverses multiples. En l'absence de renseignements, la banque est retenue dans la base de sondage. Pour éliminer les banques non valides, on s'est servi de l'annuaire du téléphone, ainsi que de divers fichiers administratifs internes.

D'après les renseignements géographiques disponibles (codes postaux), les banques de cent numéros retenues dans la base de sondage ont été regroupées pour créer des strates CA englobant, de façon aussi exacte que possible, les régions sociosanitaires. À l'intérieur de chaque strate CA, on a choisi au hasard une banque de cent numéros et généré aléatoirement un numéro compris entre 00 et 99 afin de créer un numéro de téléphone complet à 10 chiffres. Cette méthode a été répétée jusqu'à ce que l'on ait atteint le nombre requis de numéros de téléphone pour la strate CA. Comme, fréquemment, le numéro obtenu n'est pas en service ou est hors du champ d'observation, il faut générer un grand nombre de numéros supplémentaires pour atteindre la taille visée d'échantillon. Ce taux de réussite diffère selon la région. Dans le cas de l'ESCC, il variait de 15 % à 25 % parmi les 5 RSS qui y ont eu recours.

5.5 Échantillonnage des personnes interviewées

La sélection des répondants a été conçue de façon à ce que les jeunes (de 12 à 19 ans) soient surreprésentés dans l'échantillon. La stratégie d'échantillonnage adoptée tient compte des besoins des utilisateurs de données, du coût, de l'efficacité du plan d'échantillonnage, du fardeau de réponse et des contraintes opérationnelles¹⁴. La règle de sélection des personnes dans les ménages était basée sur la composition de ceux-ci en attribuant différentes probabilités de sélection aux personnes. Le tableau 5.3 fournit les probabilités de sélection des personnes du groupe d'âge 12 à 19 comparativement aux probabilités de sélection des personnes des autres groupes d'âge. À titre d'exemple, prenons un ménage de trois personnes : deux adultes et un adolescent de 15 ans. L'adolescent de 15 ans aura donc 5,8 plus de chances d'être sélectionné comparativement aux deux adultes. Il est bon de noter que pour les ménages qui se retrouvent dans une cellule « = », toutes les personnes ont la même probabilité de sélection.

¹⁴ Béland, Y., Bailie, L., Catlin, G. et Singh, M.P. CCHS and NPHS — An Improved Health Survey Program at Statistics Canada, 2000 *Proceedings of the American Statistical Association Meeting, Survey Research Methods Section*, Indianapolis: American Statistical Association, 677-682.

Tableau 5.3. Stratégie de sélection fondée sur la composition du ménage – Probabilité de sélection (12-19) comparativement à probabilité de sélection (20 ou plus)

Nombre de personnes âgées de 12 à 19 ans	Nombre de personnes âgées de 20 ans ou plus					
	0	1	2	3	4	5+
0	-	=	=	=	=	=
1	=	5,8x	4,8x	3,8x	2,8x	=
2	=	2,9x	2,4x	=	=	=
3+	=	=	=	=	=	=

5.6 Répartition de l'échantillon sur la période de collecte des données

Afin d'équilibrer la charge de travail des intervieweurs et de réduire au minimum les effets saisonniers éventuels sur des caractéristiques importantes telle que l'activité physique, dans chaque RSS, l'échantillon initial de logements/numéros de téléphone a été réparti au hasard, de façon égale, sur les 11 mois de la collecte des données (le douzième mois sert habituellement de mois de « nettoyage »). Pour commencer, chaque UPE sélectionnée au premier degré de l'échantillonnage à partir de la base aréolaire a été affectée au hasard à un trimestre de collecte des données (Q1 : janvier à mars 2003, Q2 : avril et mai 2003, Q3 : juin à août 2003 et Q4 : septembre à novembre 2003). Pour chaque trimestre de collecte, les logements sélectionnés ont alors été attribués au hasard à un mois de collecte. Pour les listes des numéros de téléphone, des échantillons indépendants ont été sélectionnés chaque mois. Cette stratégie a permis d'assurer que chaque échantillon trimestriel soit représentatif de la population canadienne faisant partie du champ d'observation de l'enquête.

5.7 Achat d'unités d'échantillonnage supplémentaires dans trois RSS de la province de Québec

Afin d'obtenir des estimations fiables à l'échelle infrarégionale, trois RSS de la province de Québec ont fourni des fonds supplémentaires afin que l'on puisse sélectionner un échantillon plus important de logements. Les unités d'échantillonnage supplémentaires ont été regroupées à l'échantillon principal afin de produire un grand fichier de données.

Les échantillons d'unités supplémentaires ont été sélectionnées entièrement à partir de la base liste de numéros de téléphone. Pour cela, on a couplé l'annuaire téléphonique du Canada à des fichiers administratifs internes afin de stratifier les numéros de téléphone publiés dans les régions infrarégionales (8 pour Québec, 7 pour les Laurentides et 8 pour l'Outaouais). Les tailles de l'échantillon sélectionné par région infraprovinciale ont été établies d'après les fonds disponibles et les exigences des RSS quant à l'obtention d'estimations fiables selon les régions infrarégionales (Québec a ajouté 1 560 unités, Laurentides a ajouté 1 630 et l'Outaouais a ajouté 1 910). Le tableau 5.4 donne la répartition des échantillons selon les régions infrarégionales.

Tableau 5.4. Répartition finale de l'échantillon, y compris les unités d'échantillonnage supplémentaires parmi les RSS de Québec, des Laurentides et de l'Outaouais

Région infraprovinciale	Taille totale (visée)
Région de Québec	2 458
Charlevoix	307
De la Jacques-Cartier	308
Haute-Ville-Des Rivières	307
La Source	307
Orléans	307
Portneuf	308
Québec-Basse-Ville-Limoilou-Vanier	307
Ste-Foy-Sillery-Laurentien	307
Région de l'Outaouais	2 528
Hull	316
Grande-Rivière	316
Gatineau	316
Pontiac	316
Domaine des Forestiers	316
Vallée-de-la-Lièvre	316
Petite-Nation	316
Les Collines-de-l'Outaouais	316
Région des Laurentides	2 363
Jean-Olivier Chénier	324
Thérèse-de-Blainville	324
Hautes-Laurentides	339
Arthur-Buies	325
Pays-d'en-Haut	383
Trois-Vallées	343
Argenteuil	325

5.8 Étude spéciale sur le mode de collecte (IPAO versus ITAO)

Dans le but de mieux comprendre les différences dans les estimations causées par le mode de collecte des données dans l'ESCC une étude spéciale a été mise en place à même le cycle 2.1 de l'ESCC. Les détails du plan d'échantillonnage sont fournis dans ce document car certains ajustements de la stratégie de pondération décrite dans le chapitre 8 réfèrent à cette étude.

L'étude sur le mode de collecte a utilisé un plan à plusieurs degrés à panel partagé pour lequel une seule base de sondage a été utilisée et où les unités secondaires d'échantillonnage ont été

randomisées (IPAO ou ITAO). L'étude a été menée dans 11 sites de façon à représenter chaque région du pays. L'étude a été réalisée durant les mois de juillet à novembre 2003. L'échantillon de chaque mode de collecte a été réparti parmi les sites de l'étude proportionnellement aux tailles d'échantillon de l'ESCC dans ces mêmes sites. Le tableau 5.5 fournit la distribution détaillée par mode de collecte des tailles d'échantillon par site.

Tableau 5.5 Tailles d'échantillon de l'étude du mode de collecte

Région sociosanitaire	IPAO	ITAO
St.John's, Terre-Neuve et Labrador	135	100
Cap Breton, Nouvelle-Écosse	125	100
Halifax, Nouvelle-Écosse	200	150
Chaudière-Appalaches, Québec	230	215
Montréal, Québec	405	390
Niagara, Ontario	235	230
Waterloo, Ontario	235	230
Winnipeg, Manitoba	320	320
Calgary, Alberta	350	290
Edmonton, Alberta	335	290
South Fraser, Colombie-Britannique	240	240
Total	2 810	2 555

L'échantillon du mode de collecte a été sélectionné à partir d'une seule base de sondage. Dans le but de minimiser l'impact sur les procédures des intervieweurs de l'équipe-terrain de l'ESCC, la base liste de numéros de téléphone a été utilisée. L'étude du mode de collecte a utilisé un plan d'échantillonnage stratifié à deux degrés où les 11 sites d'étude représentaient les strates du plan. Les unités primaires d'échantillonnage étaient les sous-divisions de recensement (SDR) alors que les numéros de téléphone étaient les unités secondaires d'échantillonnage. Dans chaque site, l'échantillon de numéros de téléphone a été sélectionné de la façon suivante :

- i. Premier degré: sélection avec probabilité proportionnelle à la taille des SDR.
- ii. Répartition de l'échantillon total (IPAO + ITAO) d'un site donné parmi les SDR sélectionnées proportionnellement à la taille des SDR.
- iii. Deuxième degré: sélection aléatoire des numéros de téléphone dans chaque SDR.

Une fois l'échantillon de numéros de téléphone obtenu, les cas pour lesquels une adresse valide n'était pas disponible ont été retirés du processus et ajoutés à l'échantillon ITAO régulier de l'ESCC. Il est à noter que ces cas n'ont pas fait partie de l'étude mais ont été intégrés à la collecte régulière de l'ESCC. Tout en contrôlant à l'échelle des SDR dans chaque site, les numéros de téléphone avec une adresse valide ont été randomisés parmi les deux modes de collecte (IPAO ou ITAO) pour ainsi constituer les deux échantillons de l'étude. Les intervieweurs de l'équipe-terrain ont alors reçu les cas de l'étude du mode de collecte (entre 20 à 60 cas) en tant que charge de travail séparée de leur charge régulière de l'ESCC car ils avaient comme instruction de n'effectuer que des interviews personnelles (IPAO) avec ces cas. (Il est bon de noter que les intervieweurs de l'équipe-terrain de l'ESCC ont l'option de compléter des interviews téléphoniques dans certaines

circonstances spéciales.) L'échantillon ITAO de l'étude du mode de collecte a simplement été ajouté aux échantillons mensuels réguliers de l'ESCC (juillet, août et septembre). (L'étude du mode de collecte était complètement transparente pour les intervieweurs des centres d'appel.) Il est bon de rappeler au lecteur que tous les cas de l'étude du mode de collecte font également partie du fichier maître de l'ESCC.

Un article¹⁵ décrivant les résultats de l'étude du mode de collecte menée à même le cycle 2.1 de l'ESCC sera publié bientôt.

¹⁵ St-Pierre, M. and Béland, Y. – Mode Effects in the Canadian Community Health Survey: a comparison of CAPI and CATI, *2004 Proceedings of the American Statistical Association Meeting, Survey Research Methods Section*, Toronto: American Statistical Association, (à être publié).

6. Collecte des données

6.1 Développement du questionnaire et méthode de collecte des données

Le questionnaire du cycle 2.1 de l' ESCC a fait l'objet d'interviews assistées par ordinateur (IAO). Des unités d'échantillonnage sélectionnées à partir de la base aréolaire, et par composition aléatoire, ont répondu aux questions suivant la méthode d'IPAO tandis que les autres unités, sélectionnées à partir des bases de sondage téléphoniques, ont répondu aux questions suivant la méthode ITAO. Dans certains cas, les intervieweurs sur le terrain ont pu compléter une partie de l'interview par téléphone.

L'IAO procure un certain nombre d'avantages quant à la qualité des données par rapport aux autres méthodes de collecte. Premièrement, le libellé des questions, comprenant les périodes de référence et les pronoms, est personnalisé automatiquement en fonction de facteurs comme l'âge et le sexe du répondant, de la date de l'interview et des réponses aux questions précédentes.

En second lieu, on applique des mesures de contrôle qui isolent les réponses incohérentes ou hors normes, et des instructions apparaissent à l'écran lorsqu'une entrée incorrecte est enregistrée. Le répondant reçoit une rétroaction immédiate et l'intervieweur peut corriger toute incohérence.

Troisièmement, le processus permet de sauter automatiquement les questions qui ne concernent pas le répondant.

6.2 Supervision et contrôle

Les intervieweurs préposés à l'ITAO ont travaillé dans les centres d'appels des bureaux centraux et ont été supervisés par un intervieweur principal travaillant dans le même bureau qu'eux. La responsabilité de la transmission des cas traités par chacun des cinq bureaux d'ITAO incombait au superviseur de projet du bureau régional, à l'intervieweur principal et à l'équipe de soutien technique.

Les intervieweurs préposés à l'IPAO ont travaillé individuellement, de leur domicile, en se servant d'un ordinateur portable et ont été supervisés à distance par les intervieweurs principaux. Ils ont transmis quotidiennement les interviews achevées au Bureau central de Statistique Canada au moyen d'une ligne téléphonique sécurisée, directement de leur domicile.

La collecte des données par ITAO n'a pas bénéficié de l'utilisation d'un ordonnanceur automatique d'appels, c'est-à-dire un système central qui optimise l'horaire des rappels et le calendrier des rendez-vous. Au lieu de cela, au début de chaque mois, un lot de cas a été affecté à chaque ordinateur personnel dans chaque bureau d'ITAO. Puis, la charge de travail sur chaque ordinateur personnel a été gérée manuellement. Cette approche s'est avérée raisonnablement efficace et le fait de ne pas avoir utilisé d'ordonnanceur d'appels n'est pas considéré comme ayant eu un effet indésirable sur la qualité des données.

6.3 Essais sur le terrain

Des essais distincts de collecte par IPAO et par ITAO ont été réalisés à la fin de l'été 2002. Les essais ont eu lieu en Alberta et au Québec.

Les principaux objectifs des essais sur la méthode IPAO étaient d'évaluer les réactions des répondants aux questions et d'obtenir des estimations quant au temps requis pour remplir les diverses sections du questionnaire. On a aussi évalué les procédures des opérations sur le terrain, la formation des intervieweurs et l'application IAO.

Les tests sur la méthode ITAO visaient des objectifs similaires. On a également évalué l'infrastructure technique des bureaux ITAO de même que les procédés d'interview propres à l'ITAO.

6.4 Techniques d'interview

Dans tous les logements choisis, on demandait à un membre du ménage bien informé de fournir l'information démographique de base sur tous les occupants du logement. Puis, on a sélectionné un membre du ménage pour une interview plus approfondie, appelée interview C2.

Les intervieweurs préposés à l'IPAO ont reçu la formation nécessaire pour procéder à une première prise de contact sur place avec chaque ménage échantillonné. Dans les cas où cette première visite s'est soldée par une non-réponse, les suivis par téléphone ont été permis. Dans le fichier de microdonnées, des indicateurs signalent si un cas a été sélectionné à partir d'une base aréolaire ou à partir d'une base téléphonique (SAMC_TYP) et si l'interview a eu lieu sur place, par téléphone ou au moyen d'une combinaison des deux techniques (ADMC_N09).

Dans les cas où le répondant sélectionné était, pour des raisons de santé physique ou mentale, incapable de répondre à l'interview, les renseignements à son sujet ont été fournis par un autre membre bien informé du ménage. Il s'agit là d'une interview par procuration. Quoique les interviewés étaient en mesure de donner des réponses exactes à la plupart des questions de l'enquête, les questions plus délicates ou personnelles allaient au-delà des connaissances d'un répondant substitut. Par conséquent, certaines questions posées dans le cadre de ces interviews par procuration sont demeurées sans réponse. Il fallait donc tout tenter pour réduire au minimum le nombre d'interviews de ce genre. La variable ADMC_PRX indique si l'interview a été réalisée par procuration ou non.

6.5 Réduction de la non-réponse

Avant même que l'intervieweur n'effectue un premier contact, les occupants du logement retenu avaient reçu une lettre de présentation et une brochure. Ces documents expliquaient l'importance de l'enquête et fournissaient des exemples sur la façon dont les données du cycle 2.1 de l'ESCC allaient être utilisées.

Les intervieweurs ont reçu comme instructions de réaliser toutes les tentatives raisonnables pour obtenir les interviews nécessaires à l'ESCC. Lorsque la visite de l'intervieweur tombait au

mauvais moment, il prenait un rendez-vous à un moment plus convenable. S'il n'y avait personne à la maison, il effectuait de nombreuses visites de rappel. Aux personnes qui refusaient dès le premier contact de participer à l'ESCC, le bureau régional envoyait une lettre insistant sur l'importance de l'enquête et de la collaboration du ménage. Suivait un second appel (ou visite) d'un intervieweur principal, d'un surveillant de projet ou un d'autre intervieweur qui tentait de convaincre les répondants de l'importance de participer à l'enquête. Au cours des derniers mois de la collecte des données, les cas de non-réponse ont été de nouveau contactés et encouragés à participer à l'enquête. Cette diligence à assurer le contact a peut-être contribué à obtenir de meilleurs résultats en maximisant le taux de réponse.

Pour pallier au problème de langue susceptible de nuire aux interviews, tous les bureaux régionaux de Statistique Canada ont embauché des intervieweurs qui parlaient plusieurs langues. Lorsqu'il le fallait, des cas étaient transférés à un intervieweur capable de remplir le questionnaire dans la langue voulue. De plus, les questions de l'enquête étaient traduites dans les langues suivantes: chinois, punjabi, inuktitut et cri.

À la fin de la collecte des données, le taux de réponse à l'échelle nationale s'est élevé à 80.7 %. Le lecteur trouvera tous les détails concernant les taux de réponse à la section 9.

6.6 Respect de la vie privée

Afin d'assurer la qualité des données recueillies, on s'est efforcé par tous les moyens de réaliser les interviews en privé. Dans certaines situations, le répondant a autorisé une autre personne à assister à l'interview. Dans le fichier de microdonnées, des indicateurs signalent si une personne autre que le répondant était présente durant l'interview (ADMC_N10) et si l'intervieweur a eu le sentiment que la présence de cette personne a influencé les réponses du répondant (ADMC_N11).

7. Traitement des données

7.1 Vérification

La vérification des données a été exécutée en grande partie par l'application d'interview assistée par ordinateur (IAO) durant la collecte des données. Les intervieweurs ne pouvaient pas entrer de valeurs hors-normes et les erreurs d'enchaînement faisaient l'objet de l'instruction de contrôle programmée « passez à ». Par exemple, l'IAO s'assurait de ne pas poser au répondant les questions non pertinentes.

En réponse à certaines données incompatibles ou inhabituelles, on a signalé des messages d'avertissement, mais sans prendre de mesures correctrices au moment de l'interview. On a plutôt mis au point, le cas échéant, des versions révisées à appliquer après la collecte des données au Bureau central. Les incohérences ont été le plus souvent corrigées en attribuant à l'une ou aux deux variables en question la valeur « non déclaré ».

7.2 Codage

On a fourni des catégories de réponses précodées pour toutes les variables appropriées. Les intervieweurs ont reçu une formation durant laquelle ils ont appris à classer les réponses recueillies dans la catégorie appropriée.

Dans les cas où la réponse donnée par le répondant ne pouvait être assignée facilement à une catégorie existante, l'intervieweur pouvait poser plusieurs questions lui permettant d'entrer une réponse en toutes lettres dans la catégorie « Autre – précisez ». Les réponses à toutes ces questions ont été examinées attentivement lors du traitement des données au Bureau central. Dans certains cas, on a donné aux réponses en toutes lettres le code d'une catégorie figurant sur la liste si la réponse faisait double emploi. On tiendra compte des réponses « Autre – précisez » fournies pour toutes les questions lors du perfectionnement des catégories de réponses en vue de futurs cycles de l'enquête.

7.3 Création de variables dérivées et groupées

Pour faciliter l'analyse des données, on a dérivé un certain nombre de variables à partir des éléments disponibles sur le questionnaire du cycle 2.1 de l'ESCC. Le cinquième caractère du nom des variables dérivées est en général un « D », « G » ou un « F ». Dans certains cas, les variables dérivées sont simples, donnant lieu à un regroupement des catégories de réponses. Dans d'autres cas, on a combiné plusieurs variables pour en créer une nouvelle. La documentation sur les variables dérivées fournit des détails sur la façon de dériver ces variables plus complexes.

7.4 Pondération

Le principe de base de l'estimation dans un échantillon aléatoire comme celui du cycle 2.1 de l'ESCC repose sur le fait que chaque personne représente, en plus d'elle-même, plusieurs autres personnes qui ne font pas partie de l'échantillon. Par exemple, dans un échantillon aléatoire simple de 2 % de la population, chaque personne en représente 50. Dans la terminologie en usage ici, nous dirons que nous avons attribué à chaque personne un facteur de pondération de 50.

L'étape de détermination des facteurs de pondération donne lieu au calcul du poids d'échantillonnage de chaque personne échantillonnée. Ce poids apparaît dans le fichier de microdonnées et doit servir à extraire des estimations de l'enquête. Par exemple, si l'on doit évaluer le nombre de personnes qui fument tous les jours, on le fait en choisissant dans l'échantillon les enregistrements des personnes qui présentent cette caractéristique et en faisant la somme des facteurs de pondération que représentent ces enregistrements.

Vous trouverez les détails sur la façon dont on calcule les poids d'échantillonnage à la section 8.

7.5 Élimination des renseignements confidentiels

Il convient de souligner que le fichier de microdonnées à grande diffusion décrit plus haut diffère, sous un nombre important d'aspects, du fichier maître de l'enquête tenu par Statistique Canada. Ces différences découlent des mesures prises pour protéger l'anonymat des répondants. La protection des répondants est assurée grâce à la suppression des valeurs individuelles, au regroupement et à l'établissement des valeurs extrêmes des variables. Les utilisateurs qui demandent l'accès à de l'information non comprise sur le fichier de microdonnées à grande diffusion, ont trois options: acheter des tableaux personnalisés, utiliser le Programme des centres de données de recherche¹⁶, ou utiliser le service de télé-accès. (Voir Section 12.3)

¹⁶ L'information la plus récente sur les Centres de données de recherche se retrouve à http://www.statcan.ca/francais/rdc/index_f.htm

8. Pondération

Pour que les estimations produites à partir de données d'enquête soient représentatives de la population couverte, et non pas seulement représentatives de l'échantillon comme tel, l'utilisateur doit incorporer les facteurs de pondération, appelés ici les poids d'enquête, dans ses calculs. Un poids d'enquête est attribué à chaque personne incluse dans l'échantillon final, c'est-à-dire dans l'échantillon de personnes ayant répondu à l'enquête. Ce poids correspond au nombre de personnes représentées par le répondant dans l'ensemble de la population.

Tel que décrit dans la section 5, l'ESCC (cycle 2.1) a eu recours à trois bases de sondage pour la sélection de son échantillon : une base aréolaire de logements agissant comme base principale, puis deux bases formées de numéros de téléphone utilisées pour compléter la base aréolaire. Puisque seulement quelques différences mineures distinguent les deux bases de numéros de téléphone pour la pondération, elles ont été traitées ensemble. On réfère à celles-ci comme faisant partie de la base téléphonique.

La stratégie de pondération a été développée en traitant séparément la base aréolaire et la base téléphonique. Les poids résultant de ces deux bases sont ensuite combinés en un seul ensemble de poids lors d'une étape appelée "intégration". Suite à quelques ajustements, ce poids intégré devient le poids final. Noter que dépendamment du besoin, une seule ou deux bases pouvaient être utilisées pour la sélection de l'échantillon dans une région sociosanitaire donnée. La stratégie de pondération s'occupe de cette particularité lors de l'étape d'intégration.

8.1 Pondération de l'échantillon

Tel que mentionné plus haut, les unités des bases aréolaire et téléphonique sont traitées séparément jusqu'à l'étape d'intégration (I1). La sous-section 8.1.1 fournit les détails de la stratégie de pondération pour la base aréolaire, puis la sous-section 8.1.2, ceux pour la base téléphonique. L'intégration des deux bases est traitée en 8.1.3, puis suivent les deux étapes finales de la pondération, c'est-à-dire l'ajustement pour contrôler la saisonnalité des données puis la poststratification, qui sont expliquées dans les sous-sections 8.1.4 et 8.1.5 respectivement.

Malgré que les deux bases aient été utilisées pour couvrir les trois territoires, certaines modifications ont dû être faites relativement à leur utilisation. Ces modifications affectent substantiellement la pondération pour ces trois régions, et celles-ci sont rapportées dans la sous-section 8.1.6.

Le diagramme A présente un sommaire des différents ajustements faisant partie de la stratégie de pondération dans l'ordre qu'ils sont appliqués. Un système de numérotation est utilisé pour identifier chaque ajustement apporté au poids et sera utilisé tout au long de la section. Les lettres *A* et *T* sont utilisées comme préfixes pour référer aux ajustements appliqués aux unités des bases Aréolaire et Téléphonique respectivement. Le préfixe *I* est quant à lui utilisé pour identifier l'ajustement d'Intégration et ceux qui suivent.

Diagramme A Sommaire de la stratégie de pondération

Base aréolaire	Base téléphonique
A0 - Poids initial	T0 - Poids initial
A1 - Accroissement de l'échantillon	T1 - Nombre de mois
A2 - Stabilisation	T2 - Retrait des unités hors champ
A3 - Retrait des unités hors champ	T3 - Couverture des bases listes
A4 - Non-réponse ménage	T4 - Combinaison des bases listes
A5 - Création du poids-personne	T5 - Non-réponse ménage
A6 - Non-réponse personne	T6 - Ménages sans téléphone
Poids final de la base aréolaire	T7 - Création du poids-personne
↗	T8 - Non-réponse personne
	T9 - Lignes multiples
	Poids final de la base téléphonique
	↘
	I1 - Intégration
	I2 - Effet saisonnier
	I3 - Poststratification
	Poids final du cycle 2.1 de l' ESCC

8.1.1 Pondération de l'échantillon provenant de la base aréolaire

A0 – Poids initial

Puisque le mécanisme utilisé pour sélectionner l'échantillon de la base aréolaire a été celui établi pour l'EPA, le poids initial a dû être calculé selon les particularités de cette enquête. D'abord, à l'intérieur de chacune des strates définies par l'EPA, des grappes (unités primaires) sont sélectionnées avec probabilités proportionnelles à la taille (selon les comptes de recensement de 1991). À l'intérieur de chacune des grappes sélectionnées, un échantillon de logements est ensuite choisi à l'aide d'un échantillonnage systématique. Le produit des probabilités de chacune de ces sélections représente la probabilité de sélection du logement et son inverse représente le poids initial du logement. Pour plus de détails sur le mécanisme de sélection, de même qu'une définition plus complète des strates et des grappes, se référer à Statistique Canada (1998)¹⁷.

A1 – Accroissement de l'échantillon

Certaines modifications ont dû être faites au mécanisme standard de l'EPA lors de la sélection de l'échantillon pour le cycle 2.1 de l'ESCC. Le plan de l'EPA peut fournir un échantillon d'environ 68 000 logements à l'échelle du pays alors que les besoins du cycle 2.1 de l' ESCC sont plus grands dans certaines régions. Les modifications apportées afin de pouvoir obtenir la taille

¹⁷ Statistique Canada (1998). *Méthodologie de l'enquête sur la population active du Canada*. Statistique Canada numéro 71-526-XPB au catalogue.

d'échantillon désirée ont été, en résumé, de répéter le même processus de sélection une deuxième fois à l'intérieur de toutes les grappes sélectionnées dans la RSS où le besoin en échantillon se faisait sentir. Ceci a eu l'effet d'accroître l'échantillon et on a dû en tenir compte dans la pondération afin de correctement représenter la probabilité de sélection. Un facteur d'ajustement représentant le taux d'accroissement de l'échantillon a donc été calculé. Cependant, cet accroissement de l'échantillon n'a pas été observé dans toutes les régions. En effet, pour certaines régions, le plan de l'EPA conduit à des tailles d'échantillon supérieures à celles requises par l'ESCC. Pour ces régions, le taux d'accroissement de l'échantillon qui est calculé représente plutôt un taux de décroissement. Le poids initial calculé en A0 est multiplié par ce facteur d'ajustement, ce qui permet d'obtenir le poids A1.

A2 – Stabilisation

Dans certaines RSS, l'accroissement de l'échantillon tel que décrit au paragraphe précédent résultait en un échantillon beaucoup plus grand que nécessaire. Une stabilisation a donc été instaurée afin de ramener la taille de l'échantillon au niveau désiré. Le processus de stabilisation consistait à sous-échantillonner des logements aléatoirement à l'intérieur de la RSS. Un facteur d'ajustement représentant l'effet de la stabilisation a donc été calculé afin de corriger la probabilité de sélection. Ce facteur, multiplié par le poids A1, produit le poids A2.

A3 – Retrait des unités hors champ

Parmi tous les logements échantillonnés, une certaine proportion de ceux-ci est, lors de la collecte, identifiée comme étant hors du champ de l'enquête. Des logements détruits ou en construction, des logements vacants, saisonniers ou secondaires, de même que des établissements, sont tous des exemples de cas hors champ pour l'ESCC. Ces logements sont tout simplement retirés de l'échantillon, ne laissant plus que les logements faisant partie du champ de l'enquête. Ces derniers conservent donc le même poids qu'à l'étape précédente que l'on appelle maintenant poids A3.

A4 – Non-réponse ménage

Lors de la collecte, une certaine proportion des ménages interviewés a inévitablement résulté en non-réponse. Ceci survient habituellement lorsque le ménage refuse de participer à l'enquête, fournit des données inutilisables, ou encore, ne peut être rejoint pour réaliser l'interview. Le poids des ménages non-répondants est redistribué aux répondants à l'aide de classes de réponse. L'algorithme CHAID (Chi-Square Automatic Interaction Detector), disponible dans Knowledge Seeker¹⁸, permet d'identifier les caractéristiques qui divisent le mieux l'échantillon en groupes selon leurs propensions à répondre. Noter que ces groupes sont formés de façon indépendante à l'intérieur de chaque RSS. Puisque l'information disponible auprès des non-répondants est très limitée, seules quelques caractéristiques telles que la période de collecte et un indicateur du milieu rural/urbain ont pu être utilisées pour la création des classes. Il s'est avéré que la caractéristique période de collecte (avec 4 périodes: janvier à mars, avril à juin, juillet à septembre, et octobre à décembre) était la plus significative pour la création des classes pour chacune des RSS.

¹⁸ ANGOS Software (1995). Knowledge Seeker IV for Windows - User's Guide. ANGOS Software International Limited.

L'indicateur rural/urbain était également une caractéristique significative pour un petit nombre de RSS. Un facteur d'ajustement a donc été calculé à l'intérieur de chaque classe de la façon suivante:

$$\frac{\text{Somme des poids A3 pour tous les ménages}}{\text{Somme des poids A3 pour tous les ménages répondants}}$$

Le poids A3 des ménages répondants a donc été multiplié par ce facteur d'ajustement pour produire le poids A4. Les ménages non-répondants sont éliminés du processus de pondération à partir de ce point.

A5 – Création du poids-personne

Puisque l'unité d'échantillonnage finale pour l'ESCC est la personne, le poids-ménage calculé jusqu'ici doit être converti en un poids-personne. Celui-ci est obtenu en multipliant le poids A4 par l'inverse de la probabilité de sélection de la personne choisie dans le ménage. Nous obtenons ainsi le poids A5. Rappelons que pour les ménages dans lesquels on retrouve un certain nombre de personnes dans les groupes d'âge 12-19 et 20+, cette probabilité de sélection de la personne est plus élevée pour les personnes du groupe 12-19 (voir section 5.5 pour plus de détails). Pour les autres ménages, cette probabilité est égale à l'inverse du nombre de personnes de 12 ans et plus dans le ménage, et ceci, peu importe la personne sélectionnée.

A6 – Non-réponse personne

Dans le cadre du cycle 2.1 de l'ESCC, une interview peut être vue comme un processus en deux étapes. Dans un premier temps, l'intervieweur obtient la liste complète des personnes vivant dans le ménage, puis par la suite interviewe la personne sélectionnée dans le ménage. Dans certains cas, les intervieweurs ne réussissent qu'à compléter la première étape, soit parce qu'ils ne peuvent entrer en contact avec la personne sélectionnée, ou encore parce que la personne sélectionnée refuse d'être interviewée. De tels cas sont définis comme étant des non-réponses à l'échelle de la personne, et un facteur d'ajustement doit être appliqué aux poids des personnes répondantes pour compenser cette non-réponse. Tout comme pour la non-réponse à l'échelle du ménage, l'ajustement est appliqué à l'intérieur de classes définies à partir des caractéristiques disponibles pour les répondants et non-répondants. Toutes les caractéristiques recueillies lors du listage des membres du ménage étaient en fait disponibles pour créer ces classes. L'algorithme CHAID a encore une fois été utilisé pour obtenir la définition des classes et le résultat final présente quelques variations dans la définition des classes d'une RSS à l'autre. Selon la RSS, les caractéristiques suivantes ont pu être utilisées pour former les classes d'ajustement : le sexe, le groupe d'âge, l'indicateur de milieu rural/urbain, le nombre de personnes dans le ménage, l'éducation, l'état matrimonial et la période de collecte. Un facteur d'ajustement est calculé à l'intérieur de chaque classe de la façon suivante:

$$\frac{\text{Somme des poids A5 pour toutes les personnes sélectionnées}}{\text{Somme des poids A5 pour toutes les personnes sélectionnées répondantes}}$$

Le poids A5 des personnes répondantes a donc été multiplié par ce facteur d'ajustement pour produire le poids A6. Les personnes non-répondantes sont éliminées de la pondération à partir de ce point.

Étant donné que cet ajustement est le dernier nécessaire pour l'échantillon provenant de la base aréolaire, le poids A6 représente donc le **poids final de la base aréolaire**. Ce poids sera plus tard intégré au poids final de la base téléphonique (section 8.1.3) pour créer le poids final du cycle 2.1 de l'ESCC.

8.1.2 Pondération de l'échantillon provenant de la base téléphonique

Tel que mentionné précédemment, la base téléphonique est en fait composée de deux bases : la base CA, puis une base liste de numéros de téléphone. Noter qu'une seule de ces deux bases peut être utilisée à l'intérieur d'une RSS. La base liste est toujours utilisée comme complément à la base aréolaire tandis que la base CA est toujours utilisée seule. Les unités provenant de ces deux bases téléphoniques sont toutefois traitées ensemble et sont donc toutes soumises aux mêmes ajustements. Il existe toutefois trois exceptions: d'abord, puisque la probabilité de sélection est relative à la base utilisée pour faire la sélection, cette probabilité sera légèrement différente dépendamment que l'unité provienne de la base CA ou de la base liste. Les autres exceptions impliquent les ajustements T3 et T4. Les détails de ces exceptions sont donnés dans les sous-sections réservées aux ajustements impliqués.

Une autre particularité propre aux unités provenant de la base téléphonique affecte la façon dont l'échantillon a été pondéré. Cette particularité concerne l'emplacement géographique des unités échantillonnées. En effet, la géographie utilisée pour sélectionner l'échantillon à partir de la base téléphonique ne répliquait pas parfaitement la géographie des RSS, ce qui a forcé certaines unités à être sélectionnées dans une certaine région alors que l'information fournie lors de l'interview les localisait plutôt dans une région avoisinante. Cette particularité a été contournée lors de la pondération en appliquant tous les ajustements relativement à la RSS assignée lors de la sélection de l'échantillon. Toutefois, puisque les unités devaient en bout de ligne appartenir à leur vraie RSS, telle qu'identifiée lors de la collecte, on a dû ajuster les poids de celles-ci comme si elles avaient fait partie de leur vraie région dès la sélection de l'échantillon. Cet ajustement a été fait via la poststratification (I3) qui est traitée plus tard dans cette section.

T0 – Poids initial

Le poids initial est calculé quelque peu différemment selon que l'échantillon provienne de la base CA ou de la base liste. Dans les deux cas, le poids initial est défini comme étant l'inverse de la probabilité de sélection, mais puisque les méthodes de sélection diffèrent, les probabilités diffèrent aussi. Pour la base CA, la sélection des numéros est faite à l'intérieur de chaque strate CA. Une strate CA représente un agrégat d'indicatifs régionaux et préfixes (IRP: les six premiers chiffres du numéro à 10 chiffres), contenant chacune des banques valides de cent numéros (voir Norris et Paton¹⁹ pour plus de détails). Conséquemment, la probabilité de sélection est le ratio

¹⁹ Norris, D.A. et Paton, D.G. (1991). L'Enquête sociale générale canadienne: bilan des cinq premières années. *Techniques d'enquête*. 17, 245-260.

entre le nombre d'unités échantillonnées et cent fois le nombre de banques présentes dans la strate CA.

Pour la base liste, les numéros de téléphone sont sélectionnés parmi tous les numéros disponibles dans la liste, et ce indépendamment pour chaque RSS. Ainsi, la probabilité de sélection correspond au ratio entre le nombre d'unités échantillonnées et le nombre de numéros de téléphone dans la liste pour la RSS. Puisque l'échantillonnage pour la base téléphonique est fait sur une base mensuelle (voir ajustement T1) et que la base liste a été mise à jour à 2 reprises durant l'enquête, le nombre de numéros disponibles dans chaque RSS a pu légèrement varié affectant ainsi la probabilité de sélection dans le temps. L'inverse de ces probabilités de sélection représente le poids initial T0.

T1 – Nombre de mois

Contrairement à la base aréolaire pour laquelle l'échantillon a été sélectionné entièrement au début du processus d'échantillonnage, des échantillons ont été tirés à chaque mois pour la base téléphonique. À chacun de ces échantillons mensuels correspond un poids initial faisant en sorte que chaque échantillon soit représentatif de la RSS. Toutefois, pour que l'échantillon total ne représente qu'une seule fois la population, un facteur d'ajustement a dû être appliqué pour réduire les poids de chaque échantillon mensuel. Le facteur d'ajustement appliqué à chaque échantillon mensuel était égal à la proportion que représentait cet échantillon mensuel parmi l'échantillon total. Or, puisque 3 versions différentes de la base liste ont été utilisées et que chaque version possède une couverture différente, l'ajustement du nombre de mois a été calculé indépendamment pour chaque version de la base liste. À partir de ce moment, l'échantillon de la base liste représente trois fois la population totale, soit une fois pour chaque version de la base liste. Les échantillons provenant des 3 bases listes sont combinés à l'étape T4 de telle sorte que la base téléphonique représente alors une seule fois la population. Le poids T1 est donc obtenu en multipliant le poids T0 par le facteur d'ajustement défini ci-dessus.

T2 - Retrait des unités hors champ

Les numéros de téléphone associés à des entreprises, des établissements ou à d'autres logements hors du champ de l'enquête, de même que les numéros hors service sont tous des exemples de cas hors champ pour la base téléphonique. Comme pour la base aréolaire, ces cas sont simplement retirés de l'échantillon, ne laissant ainsi dans l'échantillon que les logements dans le champ de l'enquête. Ces derniers conservent le même poids qu'à l'étape précédente que l'on appelle maintenant poids T2.

T3 – Couverture des bases listes

Puisque la base liste ne couvre pas certains numéros de téléphone qui sont toutefois couverts par la base CA, un ajustement doit être apporté au poids initial des unités de la base liste pour que les deux bases soient comparables en ce qui a trait à la couverture. Cet ajustement consiste à gonfler le poids des unités de la base liste proportionnellement au taux de couverture dans chaque RSS. L'estimation de ce taux de couverture a été une tâche ardue, et a pu être faite à l'aide des données recueillies auprès de l'échantillon de la base aréolaire. En effet, le questionnaire utilisé pour

l'interview des personnes sélectionnées par la base aréolaire incluait un ensemble de questions vérifiant la présence d'un téléphone dans le logement du répondant, le nombre de lignes utilisées à des fins personnelles, puis le numéro pour chacune de ces lignes. Pour dériver le taux de couverture désiré, on a simplement calculé le pourcentage des numéros de téléphone recueillis étant présents sur la base liste. Ce taux a été calculé pour chaque version de la base liste étant donné que la couverture diffère d'une version à l'autre. L'inverse de ce taux représente le facteur utilisé pour cet ajustement. Le facteur, une fois multiplié par le poids T2, produit le poids T3.

T4 - Combinaison des bases listes

L'échantillon provenant de chaque version de la base liste représente jusqu'à cette étape, la population totale des RSS où la base liste est utilisée. Les unités des trois versions doivent donc être combinées afin qu'elles ne représentent qu'une seule fois la population. Pour se faire, un facteur de combinaison tenant compte de l'importance de chaque version est calculé de la façon suivante :

$$\frac{\text{Taille d'échantillon provenant de la version de la base liste}}{\text{Taille totale d'échantillon provenant de la base liste}}$$

Ce facteur est calculé et appliqué indépendamment dans chaque RSS où la base liste a été utilisée. Pour les régions où c'est la base CA qui est utilisée, le facteur d'ajustement est égal à 1. Le poids T4 est obtenu en multipliant le poids T3 par le facteur de combinaison.

T5 - Non-réponse ménage

L'ajustement fait ici pour compenser l'effet de la non-réponse ménage est identique à celui appliqué pour la base aréolaire (ajustement A4). Comme c'était le cas pour A4, la période de collecte s'est avérée une caractéristique significative pour expliquer la non-réponse. C'est donc cette variable qui a été utilisée pour définir les classes d'ajustement. Le facteur d'ajustement calculé à l'intérieur de chaque classe a été obtenu de la façon suivante:

$$\frac{\text{Somme des poids T4 pour tous les ménages}}{\text{Somme des poids T4 pour tous les ménages répondants}}$$

Le poids T4 des ménages répondants a donc été multiplié par ce facteur d'ajustement pour produire le poids T5. Les ménages non-répondants sont éliminés à partir de ce point.

T6 - Ménages sans téléphone

Une certaine proportion de la population canadienne n'a pas accès à une ligne téléphonique résidentielle privée. Tel qu'expliqué à l'étape T3, de l'information concernant la présence d'un téléphone dans le logement du répondant est recueillie auprès de l'échantillon de la base aréolaire. Cette information a été utilisée pour estimer la proportion de ménages n'ayant pas le téléphone à l'échelle de chaque RSS. Tout comme pour T3, cette proportion est ensuite utilisée pour gonfler le poids des unités de la base téléphonique, ajustant ainsi pour la sous-représentation de la base due à

cette sous-population non couverte. Le facteur utilisé pour cet ajustement correspond à l'inverse de la proportion estimée, et une fois multiplié par le poids T5, procure le poids T6.

T7 – Création du poids-personne

Tout comme l'ajustement A5, cet ajustement permet de convertir ce qui était jusqu'à cette étape-ci un poids-ménage en un poids-personne. L'algorithme de sélection de la personne à l'intérieur du ménage étant le même que pour la base aréolaire, le calcul du facteur d'ajustement est effectué de la même façon. Ce facteur, multiplié par le poids T6, donne le poids T7.

T8 - Non-réponse personne

Cet ajustement est similaire à l'ajustement A6 utilisé pour la base aréolaire. Il consiste à compenser pour l'effet de la non-réponse à l'échelle de la personne. Tout comme pour A6, une approche par classes d'ajustement a été utilisée. Ces classes étaient définies à partir des variables disponibles pour toutes les personnes sélectionnées, répondantes ou non (voir A6 pour la liste des variables disponibles). Un facteur d'ajustement a donc été calculé à l'intérieur de chaque classe de la façon suivante:

$$\frac{\text{Somme des poids T7 pour toutes les personnes sélectionnées}}{\text{Somme des poids T7 pour toutes les personnes sélectionnées répondantes}}$$

Le poids T7 des personnes répondantes a donc été multiplié par ce facteur d'ajustement pour produire le poids T8. Les personnes non-répondantes sont éliminées à partir de ce point.

T9 - Lignes multiples

Le fait que certains ménages possèdent plus d'une ligne téléphonique résidentielle a un impact sur la pondération: plus le ménage a de lignes, meilleure est sa probabilité d'être sélectionné. Conséquemment, les poids doivent être ajustés pour tenir compte du nombre de lignes résidentielles que le ménage possède. Malgré que cette caractéristique soit relative au ménage, l'information n'est recueillie que durant l'interview auprès de la personne. C'est pour cette raison que l'ajustement est fait à ce stade-ci de la pondération. Le facteur d'ajustement représente l'inverse du nombre de lignes et le poids T9 est obtenu en multipliant ce facteur par le poids T8.

Puisque cet ajustement est le dernier nécessaire pour l'échantillon provenant de la base téléphonique, le poids T9 représente donc le **poids final de la base téléphonique**. Ce poids sera par la suite, à l'étape I1, intégré au poids final de la base aréolaire pour finalement créer le poids final du cycle 2.1 de l'ESCC.

8.1.3 Intégration des bases aréolaire et téléphonique (I1)

Cette étape consiste à intégrer les poids finaux des échantillons aréolaire et téléphonique créés jusqu'à maintenant, en un seul poids en appliquant une méthode d'intégration²⁰. Un facteur d'ajustement, compris entre 0 et 1, est déterminé de façon à représenter l'importance relative de chaque échantillon dans l'échantillon total. Cette importance relative est mesurée en termes de taille d'échantillon et d'effet de plan. Plus la proportion d'échantillon qu'une base représente dans l'échantillon total est grande, plus grande sera son importance relative dans l'échantillon total. Pour ce qui de l'effet de plan, l'importance relative sera plus grande pour les unités provenant de la base dont l'effet de plan est plus petit. Pour obtenir le facteur d'ajustement d'intégration, on calcule d'abord un facteur α , obtenu de la façon suivante:

$$\alpha = \frac{n_A}{R} \bigg/ \left(\frac{n_A}{R} + n_T \right)$$

où n_A et n_T représentent respectivement les tailles d'échantillon des bases aréolaire et téléphonique, alors que R représente le rapport médian des effets de plan estimés pour chacune des deux bases. Le poids des unités de la base aréolaire est multiplié par ce facteur α , alors que le poids des unités de la base téléphonique est multiplié par $1 - \alpha$. Noter que dans les cas où une RSS n'est couverte que par une seule base, le facteur d'ajustement est égal à 1. Le produit du facteur d'ajustement dérivé ici, par le poids final calculé auparavant (A6 ou T9 dépendant de quelle base provient l'unité), procure le poids intégré I1.

8.1.4 Effet saisonnier (I2)

L'ESCC (cycle 2.1) avait initialement planifié de répartir la collecte des données également sur les douze mois de l'année de référence de l'enquête afin de contrôler entre autres l'effet saisonnier des données recueillies. Certains événements ont toutefois affecté ce plan, de sorte qu'un ajustement additionnel a dû être ajouté pour assurer qu'il n'y ait pas d'effet saisonnier dans les estimations produites à l'aide des données du cycle 2.1 de l'ESCC. L'ajustement appliqué en I2 a été fait de façon à ce que la somme des poids des unités interviewées lors d'une des quatre saisons, représente exactement 25 % de la somme des poids de l'échantillon total. Bref, après l'application de cet ajustement, la portion d'échantillon interviewée à chaque saison représente 25 % de la population totale de chaque RSS.

Les quatre saisons définies pour l'ESCC sont les périodes couvrant septembre à novembre, décembre à février, mars à mai, puis juin à août. Le facteur d'ajustement utilisé pour contrôler l'effet saisonnier d'une personne interviewée lors de la saison S , est défini comme suit:

$$\frac{\text{Somme des poids I1 pour l'échantillon total}}{4 \times \text{somme des poids I1 de l'échantillon interviewé lors de la saison } S}$$

²⁰ Skinner, C.J. and Rao, J.N.K. (1996). Estimation in Dual Frame Surveys with Complex Designs. *Journal of the American Statistical Association*. 91, 433, 349-356.

Cet ajustement saisonnier appliqué au poids I1 permet d'obtenir le poids I2.

Noter que suite à la série d'ajustements appliqués sur les poids, il est possible que certaines unités se retrouvent avec des poids se démarquant des autres poids de leur RSS au point même de devenir aberrants. Certains répondants peuvent effectivement représenter une proportion anormalement élevée de leur RSS et ainsi influencer fortement les estimations de ces RSS. Afin d'éviter cette situation, le poids des répondants qui contribuent de façon aberrante à leur groupe RSS-âge-sexe est ajusté à la baisse selon une méthode « winsorization ».

8.1.5 Poststratification (I3)

La dernière étape nécessaire afin d'obtenir le poids final du cycle 2.1 de l' ESCC est la poststratification. La poststratification est appliquée afin d'assurer que la somme des poids finaux corresponde aux estimations de populations définies à l'échelle des RSS, pour chacun des 10 groupes d'âge-sexe d'intérêt, c'est-à-dire les cinq groupes d'âge 12-19, 20-29, 30-44, 45-64, 65+, pour chacun des deux sexes. Noter que pour l'Alberta, la poststratification a été faite en utilisant une géographie révisée contenant neuf régions au lieu des dix-sept utilisées initialement lors de la planification et du déroulement de l'enquête. Noter également que pour les trois régions du Québec qui ont fait l'achat d'échantillon supplémentaire (2403, 2407 et 2415), la poststratification a été appliquée à l'échelle de la région de CLSC plutôt que par RSS.

Les estimations de population utilisées sont basées sur les comptes du Recensement de 1996, de même que sur les comptes de naissance, décès, immigration et émigration. La moyenne des estimations mensuelles de 2003 pour chacun des croisements RSS-âge-sexe a été retenue pour réaliser la poststratification. Le poids I2 a donc été ajusté afin d'obtenir le poids final I3, à l'aide du facteur d'ajustement I3 défini comme suit:

$$\frac{\text{Estimation de population pour le groupe RSS - âge - sexe du répondant}}{\text{Somme des poids I2 pour le groupe RSS - âge - sexe du répondant}}$$

Le poids I3 correspond au ***poids final du cycle 2.1 de l' ESCC*** que l'on retrouve dans le fichier de données portant le nom de variable WTSC_M.

8.1.6 Particularités de la pondération pour les trois territoires

Tel que décrit à la section 5, le plan d'échantillonnage utilisé pour les trois territoires était quelque peu différent de celui utilisé dans les 10 provinces. La stratégie de pondération a donc dû être adaptée pour répondre à ces différences. Cette section résume les changements apportés à la stratégie expliquée aux sous-sections 8.1.1 à 8.1.5.

D'abord pour la base aréolaire, tel que mentionné à la sous-section 5.4.1, une étape additionnelle de sélection a été ajoutée pour les territoires. Chaque territoire était initialement stratifié selon des regroupements de communautés à l'intérieur desquels on a sélectionné aléatoirement une communauté. Noter que les capitales de chaque territoire formaient une strate à elles seules, et étaient donc toutes trois sélectionnées automatiquement à cette première sélection. Cette

particularité n'a eu d'effet que dans le calcul de la probabilité de sélection, et donc dans la valeur du poids initial (A0). Une fois ce poids initial calculé, la même série d'ajustements (A1 à A6) a été appliquée aux unités de la base aréolaire. Les classes d'ajustement pour les non-réponses ménage et personne ont été construites à l'aide du même ensemble de variables disponibles pour les provinces.

Pour ce qui est de la pondération des unités de la base téléphonique, mentionnons tout d'abord que seule la base CA a été utilisée, et ce, uniquement à l'intérieur des capitales du Yukon et des Territoires du Nord-Ouest. Ceci élimine donc le besoin d'avoir recours aux ajustements T3 (couverture des bases listes) et T4 (combinaison des bases listes). Les autres ajustements de la base téléphonique ont tous été appliqués. Finalement, l'ajustement T6 (ménages sans téléphone) a aussi subi une légère modification puisque la base CA était utilisée uniquement dans les capitales. Les proportions de ménages sans téléphone ont été dérivées, tout comme pour les provinces, à partir des données de la base aréolaire, mais en excluant toutefois du calcul les données des ménages situés à l'extérieur des capitales.

Les deux ensembles de poids (aréolaire et téléphonique) ont ensuite été intégrés, puis ajustés pour la saisonnalité et finalement poststratifiés de façon semblable à ce qui a été fait pour les provinces, à l'exception de deux détails. D'abord, l'intégration a été appliquée uniquement pour les unités situées dans les capitales du Yukon et des Territoires du Nord-Ouest; les autres communautés ayant été couvertes uniquement par la base aréolaire. Le second détail a trait à la saisonnalité. Étant donné qu'une forte concentration des interviews a été menée sur une très courte période de temps dans le territoire du Nunavut, l'ajustement pour la saisonnalité n'a pu être appliqué de façon efficace. Les estimations produites pour le Nunavut à partir de ces poids ne tiendront donc pas compte d'un possible effet saisonnier des données.

8.1.7 Particularités de la pondération pour les RSS où a eu lieu l'étude du mode de collecte

Un traitement différent est réservé aux unités provenant des 11 régions où a eu lieu l'étude sur l'effet du mode de collecte. Pour ces régions, l'échantillon provient de trois sources : base aréolaire, base téléphonique régulière et étude sur l'effet du mode de collecte. Le tout doit être intégré pour former un seul échantillon pour la région. Noter que le poids spécifiquement calculé pour l'étude de l'effet du mode de collecte n'est pas utilisable dans le cas présent puisque les paramètres de l'ESCC sont très différents de ceux de l'étude.

Puisque les cas de l'étude sur l'effet du mode de collecte proviennent de la base liste, ils sont traités avec les cas de la base téléphonique. Cependant, certains ajustements doivent être apportés au processus régulier de pondération de la base téléphonique. En fait, les unités de la base téléphonique sont traitées séparément des unités de l'étude jusqu'à ce qu'on intègre les deux échantillons. Cette intégration est effectuée juste avant le traitement de la non-réponse ménage de la base téléphonique. La présente section décrit les différences qui sont apportées à la pondération des unités de la base téléphonique pour les 11 régions où a eu lieu l'étude de l'effet du mode de collecte.

T0 – Poids initial

Contrairement à l'échantillon régulier provenant de la base téléphonique, les numéros de téléphone sélectionnés pour faire partie de l'étude n'ont pas été sélectionnés suivant un plan aléatoire simple à l'intérieur d'une RSS. Un échantillon de subdivisions de recensement (SDR) a d'abord été sélectionné et parmi les SDR choisies, un échantillon de numéros de téléphone a été sélectionné. Ce degré d'échantillonnage supplémentaire a été ajouté pour éviter qu'un intervieweur menant des interviews en personne ait à se déplacer partout dans la RSS. Cette particularité associée aux cas de l'étude est donc incorporée dans le calcul du poids initial.

T1 – Nombre de mois

L'ajustement pour le nombre de mois n'est pas appliqué aux unités provenant de l'étude sur l'effet du mode de collecte. Contrairement à l'échantillon provenant de la base téléphonique qui est sélectionné à chaque mois, un seul échantillon a été sélectionné pour l'étude. L'ajustement pour le nombre de mois devient donc inutile pour les cas de l'étude. De plus, puisque l'échantillon de l'étude remplace l'échantillon qui était initialement prévu pour la base téléphonique, le nombre de mois pour lequel l'échantillon provient de la base liste passe de 11 à 8 pour ces RSS.

T2 – Retrait des unités hors champ

Pour les cas faisant partie de l'étude sur l'effet du mode de collecte, le concept d'unités hors champ diffère selon le mode d'interview impliqué. En effet, les intervieweurs pour la composante téléphonique composent simplement le numéro de téléphone sélectionné et si le numéro conduit à un cas dans le champ de l'enquête, ils procèdent à l'interview. De leur côté, les intervieweurs qui réalisent les interviews en personne se rendent à l'adresse qui était associée au numéro de téléphone sélectionné au moment où la base liste a été créée. Ils effectuent l'interview si le logement est dans le champ de l'enquête peu importe si le numéro de téléphone du ménage présent à l'adresse est le même que celui qui avait été choisi. Il est donc possible que le numéro de téléphone qui avait été sélectionné soit maintenant hors service mais qu'il y ait quand même un ménage dans le champ de l'enquête à l'adresse associée. Si toutes les interviews en personne avaient été faites au téléphone, il y aurait donc eu plus de cas hors champ et la perte de poids due au retrait des unités hors champ aurait été plus importante. Cette faible perte de poids due au retrait des hors champ pour les interviews en personne fait en sorte qu'on se retrouve avec un surplus de poids et ainsi une surestimation de la couverture réelle de la base téléphonique. Un ajustement est appliqué aux unités pour lesquelles l'interview a été faite en personne faisant en sorte que la perte de poids (en proportion) est la même pour ces unités que pour les unités de la composante téléphonique à l'intérieur d'une RSS. De cette façon, on s'assure de mesurer la couverture réelle de la base téléphonique au moment de sa création.

T3 – Couverture des bases listes

Aucun changement n'est apporté à cet ajustement. Il est appliqué autant aux cas faisant partie de l'étude que ceux de l'échantillon régulier.

T4 - Combinaison des bases listes

Cet ajustement s'applique uniquement aux unités qui proviennent de l'échantillon régulier de la base téléphonique. Il n'est pas applicable aux cas qui font partie de l'étude puisqu'ils proviennent tous de la même version de la base liste.

T4b – Intégration de l'étude sur le mode de collecte

Jusqu'à cette étape, les cas de l'échantillon régulier de la base téléphonique étaient traités séparément des cas de l'étude sur l'effet du mode de collecte. Cette étape consiste à combiner les poids de ces deux composantes en un seul poids. Afin d'intégrer les deux échantillons, l'ajustement suivant est calculé pour chaque composante (échantillon régulier et étude) à l'intérieur de chaque RSS:

$$\frac{\text{Taille d' échantillon pour la composante}}{\text{Taille d' échantillon totale pour les 2 composantes}}$$

Ce nouvel ajustement, multiplié par le poids T4, produit le poids T4b. À partir de cette étape, la pondération des régions faisant partie de l'étude sur le mode de collecte ne diffère pas de celle des autres régions. Les ajustements suivants sont donc appliqués de la même façon que décrite à la section 8.1.2.

9. Qualité des données

9.1 Taux de réponse

Au total et après avoir retiré les unités hors du champ de l'enquête, 166 222 ménages ont été sélectionnés pour participer à l'ESCC (cycle 2.1). De ce nombre, 144 836 ont accepté de participer à l'enquête ce qui résulte en un taux de réponse à l'échelle du ménage de 87,1 %. Parmi ces ménages répondants, 144 836 personnes ont été sélectionnées (une personne par ménage) pour participer à l'enquête parmi lesquelles 134 072 ont accepté de le faire ce qui résulte en un taux de réponse à l'échelle de la personne de 92,6 %. À l'échelle canadienne, un taux de réponse combiné de 80,7 % a donc été observé à l'ESCC (cycle 2.1). Le tableau 9.1 donne les taux de réponse combinés ainsi que l'information pertinente au calcul de ceux-ci pour chaque région sociosanitaire ou regroupement de régions sociosanitaires.

Le plan d'échantillonnage du cycle 2.1 de l'ESCC en Alberta a été conçu en juillet 2002 en utilisant les limites géographiques des 17 régions sociosanitaires en vigueur à ce moment-là. L'année suivante, soit le 1^{er} avril 2003, le gouvernement de l'Alberta redéfinissait les limites de leurs régions sociosanitaires. Les limites pour les 9 nouvelles régions se retrouvent sur le fichier de microdonnées à grande diffusion. Il n'est toutefois pas approprié de diffuser les taux de réponse pour ces nouvelles régions.

On décrit dans ce qui suit de quelle façon les différentes composantes de l'équation doivent être manipulées afin de calculer correctement les taux de réponse combinés.

Taux de réponse à l'échelle du ménage

$$\text{HHRR} = \frac{\text{\# de ménages répondants provenant des 2 bases}}{\text{tous les ménages faisant partie du champ de l'enquête provenant des 2 bases}}$$

Taux de réponse à l'échelle de la personne

$$\text{PPRR} = \frac{\text{\# de répondants provenant des 2 bases}}{\text{toutes les personnes sélectionnées provenant des 2 bases}}$$

$$\text{Taux de réponse combiné} = \text{HHRR} \times \text{PPRR}$$

Voici maintenant un exemple de calcul du taux de réponse combiné pour le Canada en utilisant l'information fournie dans le tableau 9.1.

$$\text{HHRR} = \frac{68\,966 + 75\,870}{77\,528 + 88\,694} = \frac{144\,836}{166\,222} = 0,871$$

$$\text{PPRR} = \frac{64\,656 + 69\,416}{68\,966 + 75\,870} = \frac{134\,072}{144\,836} = 0,926$$

$$\begin{aligned}\text{Taux de réponse combiné} &= 0,871 \times 0,926 \\ &= 0,807 \\ &= \mathbf{80,7 \%}.\end{aligned}$$

Table 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases /
Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							Tous les cas
Prov.	Health Region	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp.	Pers. resp. rates	Resp. rates	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp.	Pers. resp. rates	Resp. rates	Combined resp. rates
	Région socio-sanitaire	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	Taux de rép. combiné
CA	Total	77528	68966	89.0	68966	64656	93.8	83.4	88694	75870	85.5	75870	69416	91.5	78.3	80.7
NL	Total	2357	2204	93.5	2204	2066	93.7	87.7	2303	2140	92.9	2140	1988	92.9	86.3	87.0
	10901	484	433	89.5	433	401	92.6	82.9	496	462	93.1	462	427	92.4	86.1	84.5
	10902	508	487	95.9	487	457	93.8	90.0	448	415	92.6	415	387	93.3	86.4	88.3
	10903	487	466	95.7	466	431	92.5	88.5	378	352	93.1	352	328	93.2	86.8	87.7
	10904*	878	818	93.2	818	777	95.0	88.5	981	911	92.9	911	846	92.9	86.2	87.3
PE	Total	1120	1024	91.4	1024	970	94.7	86.6	1352	1189	87.9	1189	1092	91.8	80.8	83.4
	11901*	677	627	92.6	627	601	95.9	88.8	846	742	87.7	742	678	91.4	80.1	84.0
	11903	443	397	89.6	397	369	92.9	83.3	506	447	88.3	447	414	92.6	81.8	82.5
NS	Total	2799	2548	91.0	2548	2392	93.9	85.5	3092	2751	89.0	2751	2564	93.2	82.9	84.1
	12901	432	411	95.1	411	388	94.4	89.8	448	391	87.3	391	369	94.4	82.4	86.0
	12902	309	284	91.9	284	271	95.4	87.7	427	392	91.8	392	361	92.1	84.5	85.9
	12903	383	367	95.8	367	350	95.4	91.4	399	372	93.2	372	343	92.2	86.0	88.6
	12904	440	397	90.2	397	377	95.0	85.7	399	345	86.5	345	327	94.8	82.0	83.9
	12905	484	440	90.9	440	400	90.9	82.6	542	477	88.0	477	446	93.5	82.3	82.5
	12906	751	649	86.4	649	606	93.4	80.7	877	774	88.3	774	718	92.8	81.9	81.3
NB	Total	2909	2675	92.0	2675	2517	94.1	86.5	2801	2548	91.0	2548	2412	94.7	86.1	86.3
	13901	514	469	91.2	469	439	93.6	85.4	480	441	91.9	441	412	93.4	85.8	85.6
	13902	522	488	93.5	488	463	94.9	88.7	479	425	88.7	425	397	93.4	82.9	85.9
	13903	572	515	90.0	515	490	95.1	85.7	445	416	93.5	416	390	93.8	87.6	86.5
	13904*	609	561	92.1	561	527	93.9	86.5	697	636	91.2	636	610	95.9	87.5	87.1
	13906*	692	642	92.8	642	598	93.1	86.4	700	630	90.0	630	603	95.7	86.1	86.3
QC	Total	14381	12663	88.1	12663	11874	93.8	82.6	21023	17118	81.4	17118	15725	91.9	74.8	78.0
	24901	776	717	92.4	717	698	97.4	89.9	612	536	87.6	536	500	93.3	81.7	86.3
	24902	840	778	92.6	778	739	95.0	88.0	756	676	89.4	676	635	93.9	84.0	86.1
	24903	1078	927	86.0	927	880	94.9	81.6	3621	2849	78.7	2849	2598	91.2	71.7	74.0
	24904	933	827	88.6	827	795	96.1	85.2	979	821	83.9	821	753	91.7	76.9	81.0
	24905	746	646	86.6	646	583	90.2	78.2	701	608	86.7	608	565	92.9	80.6	79.3
	24906	2117	1751	82.7	1751	1637	93.5	77.3	1727	1257	72.8	1257	1131	90.0	65.5	72.0
	24907	780	708	90.8	708	656	92.7	84.1	3474	2860	82.3	2860	2606	91.1	75.0	76.7
	24908	669	604	90.3	604	561	92.9	83.9	622	542	87.1	542	507	93.5	81.5	82.7
	24909	714	625	87.5	625	580	92.8	81.2	645	553	85.7	553	510	92.2	79.1	80.2
	24911	707	655	92.6	655	628	95.9	88.8	532	446	83.8	446	389	87.2	73.1	82.1
	24912	828	745	90.0	745	717	96.2	86.6	837	704	84.1	704	655	93.0	78.3	82.4
	24913	940	803	85.4	803	681	84.8	72.4	791	626	79.1	626	576	92.0	72.8	72.6
	24914	882	780	88.4	780	721	92.4	81.7	813	673	82.8	673	629	93.5	77.4	79.6
	24915	888	777	87.5	777	734	94.5	82.7	3327	2652	79.7	2652	2448	92.3	73.6	75.5

Table 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases /
Tableau 9.1																Tous les cas
Prov.	Health Region	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp.	Pers. resp. rates	Resp. rates	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp.	Pers. resp. rates	Resp. rates	Combined resp. rates
	Région socio-sanitaire	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	Taux de rép. combiné
	24916	1483	1320	89.0	1320	1264	95.8	85.2	1586	1315	82.9	1315	1223	93.0	77.1	81.0
ON	Total	24760	21749	87.8	21749	20326	93.5	82.1	29701	25066	84.4	25066	22451	89.6	75.6	78.5
	35926	520	424	81.5	424	391	92.2	75.2	613	519	84.7	519	471	90.8	76.8	76.1
	35927	553	481	87.0	481	442	91.9	79.9	532	449	84.4	449	412	91.8	77.4	78.7
	35930	1057	912	86.3	912	841	92.2	79.6	1146	984	85.9	984	874	88.8	76.3	77.8
	35931	378	314	83.1	314	295	93.9	78.0	530	458	86.4	458	409	89.3	77.2	77.5
	35933	498	444	89.2	444	426	95.9	85.5	622	529	85.0	529	491	92.8	78.9	81.9
	35934	467	422	90.4	422	401	95.0	85.9	493	417	84.6	417	373	89.4	75.7	80.6
	35935	511	420	82.2	420	398	94.8	77.9	588	483	82.1	483	432	89.4	73.5	75.5
	35936	843	770	91.3	770	724	94.0	85.9	902	761	84.4	761	684	89.9	75.8	80.7
	35937	1111	894	80.5	894	811	90.7	73.0	1158	976	84.3	976	852	87.3	73.6	73.3
	35938	302	287	95.0	287	278	96.9	92.1	871	725	83.2	725	657	90.6	75.4	79.7
	35939*	391	357	91.3	357	332	93.0	84.9	1199	1030	85.9	1030	945	91.7	78.8	80.3
	35940	424	408	96.2	408	397	97.3	93.6	479	404	84.3	404	378	93.6	78.9	85.8
	35941	586	510	87.0	510	492	96.5	84.0	651	560	86.0	560	504	90.0	77.4	80.5
	35942	501	465	92.8	465	440	94.6	87.8	554	472	85.2	472	433	91.7	78.2	82.7
	35943	586	502	85.7	502	468	93.2	79.9	611	503	82.3	503	450	89.5	73.6	76.7
	35944	984	856	87.0	856	816	95.3	82.9	994	855	86.0	855	781	91.3	78.6	80.7
	35945	299	261	87.3	261	244	93.5	81.6	697	582	83.5	582	533	91.6	76.5	78.0
	35946	941	823	87.5	823	771	93.7	81.9	1101	926	84.1	926	831	89.7	75.5	78.5
	35947*	636	590	92.8	590	557	94.4	87.6	934	818	87.6	818	758	92.7	81.2	83.8
	35949	266	234	88.0	234	207	88.5	77.8	598	508	84.9	508	456	89.8	76.3	76.7
	35951	1212	1089	89.9	1089	1017	93.4	83.9	1334	1131	84.8	1131	1030	91.1	77.2	80.4
	35952	470	425	90.4	425	387	91.1	82.3	485	432	89.1	432	386	89.4	79.6	80.9
	35953	1521	1325	87.1	1325	1233	93.1	81.1	1503	1244	82.8	1244	1057	85.0	70.3	75.7
	35955	200	171	85.5	171	166	97.1	83.0	889	761	85.6	761	688	90.4	77.4	78.4
	35956	490	448	91.4	448	431	96.2	88.0	494	433	87.7	433	397	91.7	80.4	84.1
	35957	476	425	89.3	425	394	92.7	82.8	477	403	84.5	403	367	91.1	76.9	79.9
	35958	572	511	89.3	511	472	92.4	82.5	687	592	86.2	592	536	90.5	78.0	80.1
	35960	761	705	92.6	705	664	94.2	87.3	1012	866	85.6	866	784	90.5	77.5	81.7
	35961	640	591	92.3	591	543	91.9	84.8	676	589	87.1	589	519	88.1	76.8	80.7
	35962	540	497	92.0	497	479	96.4	88.7	635	548	86.3	548	488	89.1	76.9	82.3
	35965	934	812	86.9	812	757	93.2	81.0	1087	921	84.7	921	836	90.8	76.9	78.8
	35966	613	571	93.1	571	542	94.9	88.4	750	647	86.3	647	581	89.8	77.5	82.4
	35968	890	763	85.7	763	726	95.2	81.6	966	788	81.6	788	680	86.3	70.4	75.8
	35970	1096	980	89.4	980	928	94.7	84.7	1217	996	81.8	996	865	86.8	71.1	77.5
	35995	2491	2062	82.8	2062	1856	90.0	74.5	2216	1756	79.2	1756	1513	86.2	68.3	71.6

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
Prov.	Health Region	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp.	Pers. resp. rates	Resp. rates	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp.	Pers. resp. rates	Resp. rates	Combined resp. rates
	Région socio-sanitaire	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	Taux de rép. combiné
MB	Total	4137	3786	91.5	3786	3617	95.5	87.4	4810	4331	90.0	4331	4015	92.7	83.5	85.3
	46910	1242	1080	87.0	1080	1033	95.6	83.2	1478	1320	89.3	1320	1210	91.7	81.9	82.5
	46915*	732	681	93.0	681	650	95.4	88.8	865	789	91.2	789	744	94.3	86.0	87.3
	46920*	516	490	95.0	490	469	95.7	90.9	627	559	89.2	559	515	92.1	82.1	86.1
	46930	356	332	93.3	332	322	97.0	90.4	438	405	92.5	405	377	93.1	86.1	88.0
	46940	432	406	94.0	406	379	93.3	87.7	498	450	90.4	450	423	94.0	84.9	86.2
	46960*	859	797	92.8	797	764	95.9	88.9	904	808	89.4	808	746	92.3	82.5	85.6
SK	Total	4530	4104	90.6	4104	3887	94.7	85.8	4467	3965	88.8	3965	3700	93.3	82.8	84.3
	47901*	1140	1050	92.1	1050	973	92.7	85.4	945	846	89.5	846	793	93.7	83.9	84.7
	47904	740	652	88.1	652	620	95.1	83.8	678	598	88.2	598	554	92.6	81.7	82.8
	47905*	702	670	95.4	670	659	98.4	93.9	627	545	86.9	545	507	93.0	80.9	87.7
	47906	787	698	88.7	698	662	94.8	84.1	764	680	89.0	680	639	94.0	83.6	83.9
	47907*	720	662	91.9	662	622	94.0	86.4	582	520	89.3	520	489	94.0	84.0	85.3
	47909*	441	372	84.4	372	351	94.4	79.6	871	776	89.1	776	718	92.5	82.4	81.5
AB	Total	7930	7165	90.4	7165	6686	93.3	84.3	8846	7805	88.2	7805	7185	92.1	81.2	82.7
	48901	566	515	91.0	515	492	95.5	86.9	485	436	89.9	436	400	91.7	82.5	84.9
	48902	397	355	89.4	355	335	94.4	84.4	460	424	92.2	424	381	89.9	82.8	83.5
	48903*	675	612	90.7	612	576	94.1	85.3	708	628	88.7	628	584	93.0	82.5	83.9
	48904	1362	1209	88.8	1209	1135	93.9	83.3	1349	1149	85.2	1149	1063	92.5	78.8	81.1
	48906	577	531	92.0	531	487	91.7	84.4	543	469	86.4	469	436	93.0	80.3	82.4
	48907	460	437	95.0	437	419	95.9	91.1	464	409	88.1	409	382	93.4	82.3	86.7
	48908*	672	609	90.6	609	556	91.3	82.7	809	729	90.1	729	676	92.7	83.6	83.2
	48910	1252	1096	87.5	1096	1050	95.8	83.9	1250	1079	86.3	1079	990	91.8	79.2	81.5
	48911	414	378	91.3	378	351	92.9	84.8	453	397	87.6	397	371	93.5	81.9	83.3
	48912	570	516	90.5	516	495	95.9	86.8	401	361	90.0	361	335	92.8	83.5	85.5
	48913	440	403	91.6	403	319	79.2	72.5	366	332	90.7	332	299	90.1	81.7	76.7
	48914*	545	504	92.5	504	471	93.5	86.4	1558	1392	89.3	1392	1268	91.1	81.4	82.7
BC	Total	9713	8518	87.7	8518	7959	93.4	81.9	10056	8757	87.1	8757	8099	92.5	80.5	81.2
	59911	231	224	97.0	224	218	97.3	94.4	473	432	91.3	432	409	94.7	86.5	89.1
	59912	311	290	93.2	290	282	97.2	90.7	397	359	90.4	359	345	96.1	86.9	88.6
	59913	784	720	91.8	720	687	95.4	87.6	639	549	85.9	549	518	94.4	81.1	84.7
	59914	616	565	91.7	565	547	96.8	88.8	523	468	89.5	468	441	94.2	84.3	86.7
	59921	640	587	91.7	587	549	93.5	85.8	602	505	83.9	505	469	92.9	77.9	82.0
	59922	999	870	87.1	870	821	94.4	82.2	866	752	86.8	752	692	92.0	79.9	81.1
	59923	1004	886	88.2	886	776	87.6	77.3	996	827	83.0	827	754	91.2	75.7	76.5
	59931	397	354	89.2	354	333	94.1	83.9	625	551	88.2	551	496	90.0	79.4	81.1
	59932	1061	854	80.5	854	808	94.6	76.2	921	788	85.6	788	711	90.2	77.2	76.6

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
Prov.	Health Region	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp. pers.	Pers. resp. rates	Resp. rates	# in scope HH	# resp. HH	HH resp. rates	# pers. select.	# resp. pers.	Pers. resp. rates	Resp. rates	Combined resp. rates
	Région socio- sanitaire	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	# mén. cibles	# mén. rép.	Taux de rép. mén.	# pers. sélect.	# rép.	Taux de rép. pers.	Taux de rép.	Taux de rép. combiné
	59933	689	604	87.7	604	547	90.6	79.4	600	522	87.0	522	475	91.0	79.2	79.3
	59941	852	742	87.1	742	692	93.3	81.2	702	585	83.3	585	538	92.0	76.6	79.2
	59942	562	472	84.0	472	433	91.7	77.0	539	477	88.5	477	437	91.6	81.1	79.0
	59943	320	310	96.9	310	302	97.4	94.4	628	580	92.4	580	553	95.3	88.1	90.2
	59951*	716	588	82.1	588	540	91.8	75.4	1045	935	89.5	935	859	91.9	82.2	79.4
	59952	531	452	85.1	452	424	93.8	79.8	500	427	85.4	427	402	94.1	80.4	80.1
Terr.	60901*	2892	2530	87.5	2530	2362	93.4	81.7	243	200	82.3	200	185	92.5	76.1	81.2

* = régions sociosanitaires regroupées

9.2 Erreurs dans les enquêtes

L'enquête permet de produire des estimations fondées sur l'information recueillie à partir d'un échantillon de personnes. On aurait pu obtenir des estimations quelque peu différentes si on avait effectué 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 l'erreur due à l'échantillonnage des estimations.

Les erreurs qui ne sont pas liées à l'échantillonnage peuvent être commises à presque toutes les étapes d'une enquête. Il est possible que les intervieweurs comprennent mal les instructions, que les répondants fassent des erreurs en complétant le questionnaire, que les réponses soient mal saisies et que des erreurs se produisent au moment du traitement et de la totalisation des données. Tous ces exemples représentent des erreurs non dues à l'échantillonnage.

9.2.1 Erreurs non dues à l'échantillonnage

Sur un grand nombre d'observations, les erreurs aléatoires auront peu d'effet sur les estimations tirées de l'enquête. Toutefois, les erreurs qui se produisent systématiquement contribueront à des biais dans les estimations de l'enquête. On a consacré beaucoup de temps et d'efforts à réduire les erreurs non dues à l'échantillonnage dans l'enquête. 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, une formation poussée sur les méthodes d'enquête et le questionnaire et l'observation des intervieweurs afin de déceler les problèmes. La mise à l'essai de l'application IAO et les essais sur le terrain ont également été au nombre des procédures essentielles pour réduire au maximum les erreurs de collecte de données.

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 de non-réponse partielle (le fait de ne pas répondre à une ou plusieurs questions) à une non-réponse totale. Dans le cas du cycle 2.1 de l'ESCC, il n'y a presque pas eu de non-réponse partielle car une fois le questionnaire débuté les répondants avaient tendance à le terminer. Il y a eu non-réponse totale lorsque la personne sélectionnée pour participer à l'enquête a refusé de le faire ou que l'intervieweur a été incapable d'entrer en contact avec elle. 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 ceux qui n'ont pas répondu. Voir la section 8 pour avoir de plus amples détails sur la correction de la pondération pour la non-réponse.

9.2.2 Erreurs dues à l'échantillonnage

Étant donné que les estimations d'une enquête par sondage comportent inévitablement des erreurs dues à l'échantillonnage, de bonnes méthodes statistiques exigent que les chercheurs fournissent aux utilisateurs une certaine indication de l'ampleur de cette erreur. La mesure de l'importance éventuelle des erreurs dues à l'échantillonnage est fondée sur l'écart type des estimations tirées

des résultats de l'enquête. Cependant, en raison de la grande 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 (CV), s'obtient en divisant l'écart type de l'estimation par l'estimation elle-même et on l'exprime en pourcentage de l'estimation.

Par exemple, supposons qu'une personne estime que 25 % des Canadiens âgés de 12 ans et plus sont des fumeurs réguliers et que cette estimation comporte un écart type de 0,003. On calcule alors le CV de cette estimation de la façon suivante :

$$(0,003/0,25) \times 100 \% = 1,20 \%$$

Statistique Canada utilise fréquemment les résultats du CV pour l'analyse des données et conseille vivement aux utilisateurs produisant des estimations à partir des fichiers de données du cycle 2.1 de l'ESCC de faire de même. Pour plus d'information sur le calcul des CV, voir la section 11. Pour consulter les lignes directrices sur la façon d'interpréter les résultats du CV, se référer au tableau à la fin de la sous-section 10.4.

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 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 ce fichier de microdonnées en vue d'être publiées ou diffusées de toute autre façon correspondent à celles produites par Statistique Canada, 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, il est vivement conseillé à l'utilisateur d'indiquer la raison de ces divergences dans le ou les documents à publier ou à diffuser.

f) Des estimations non arrondies ne doivent être publiées ou diffusées de toute autre façon en aucune circonstance. Des estimations non arrondies donnent l'impression d'être beaucoup plus précises qu'elles ne le sont en réalité.

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

Le plan d'échantillonnage utilisé pour cette enquête 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 de Statistique Canada.

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

10.2.1 Définitions des catégories d'estimations : de type nominal par opposition à quantitatives

Avant d'exposer la façon de totaliser et d'analyser les données de l'enquête, 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 rentrent dans 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 :

Actuellement, est-ce que ... fume(z) des cigarettes tous les jours, à l'occasion ou jamais?
(SMKC_202)

- ☐ Tous les jours
- ☐ À l'occasion
- ☐ 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 :

Actuellement, combien de cigarettes est-ce que ... fume(z) chaque jour?
(**SMKC_204**)

|_|_| 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 :

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

10.2.3 Totalisation d'estimations quantitatives

Pour obtenir l'estimation d'une somme ou d'une moyenne pour une variable quantitative, on procède aux étapes suivantes (seule l'étape a) est nécessaire pour obtenir l'estimation pour une somme) :

- 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});
- faire la somme des poids finals des enregistrements contenant la variable étudiée pour obtenir le dénominateur (\hat{Y});
- 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 calcule d'abord le numérateur (\hat{X}) en sommant le produit de la variable **SMKC_204** par le poids, **WTSC_M** pour tous les enregistrements pour lesquels la valeur

de la variable **SMKC_202** est « tous les jours ». On obtient ensuite le dénominateur (\hat{Y}) en additionnant le poids final de tous les enregistrements pour lesquels la valeur de la variable **SMKC_202** est « tous les jours ». Le nombre moyen de cigarettes fumées chaque jour par les personnes qui fument tous les jours est finalement obtenu en divisant (\hat{X}) par (\hat{Y}).

10.3 Lignes directrices pour l'analyse statistique

L'ESCC se fonde sur un plan de sondage complexe qui prévoit une stratification et un échantillonnage à plusieurs degrés, ainsi que 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), une méthode permet de corriger les résultats obtenus des progiciels courants de façon à ce qu'il soit plus adéquat. Cette méthode consiste à rééchelonner 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 utilisant dans l'analyse un poids égal au poids original divisé par la moyenne des poids originaux pour les unités échantillonnées (personnes) qui contribuent à l'estimation en question.

Pour permettre à l'utilisateur d'évaluer la qualité des totalisations estimées d'après les données, Statistique Canada a produit un ensemble de tableaux de variabilité d'échantillonnage approximative (couramment appelées « Tableaux des CV ») pour l'ESCC. On peut employer ces tableaux pour obtenir des coefficients de variation approximatifs pour les estimations de type nominal et les proportions. Pour plus de détails, consulter la section 11.

10.4 Lignes directrices pour la diffusion

Avant de diffuser et/ou de publier des estimations tirées des fichiers de microdonnées, l'utilisateur doit d'abord déterminer le nombre de répondants dans l'échantillon ayant la caractéristique à l'étude (par exemple, le nombre de répondants qui fument lorsqu'on s'intéresse à la proportion de fumeurs pour une population donnée). 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 arrondie et suivre les lignes directrices qui suivent.

Table 10.1 : Lignes directrices relatives à la variabilité d'échantillonnage

Type d'estimation	c.v. (en %)	Lignes directrices
Acceptable	$0,0 \leq \text{c.v.} \leq 16,6$	On peut envisager une diffusion générale non restreinte des estimations. Aucune annotation particulière n'est nécessaire.
Marginale	$16,6 < \text{c.v.} \leq 33,3$	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	$\text{c.v.} > 33,3$	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. »

11. Tableaux de la variabilité d'échantillonnage approximative

Afin de permettre aux utilisateurs d'avoir facilement accès à des coefficients de variation qui s'appliqueraient à une multitude d'estimations de type nominal obtenues à partir de ce fichier de microdonnées, Statistique Canada a produit un ensemble de tableaux de la variabilité d'échantillonnage approximative. Ces tableaux permettent aux utilisateurs d'obtenir un coefficient de variation approximatif selon la taille de l'estimation calculée à partir des données de l'enquête.

Les coefficients de variation (CV) dans ces tableaux sont calculés en employant la formule de la variance utilisée pour l'échantillonnage aléatoire simple et en y incorporant un facteur qui reflète la structure en grappes à plusieurs degrés du plan d'échantillonnage. Pour obtenir ce facteur, appelé *effet du plan*, on a d'abord calculé les effets du plan pour une vaste gamme de caractéristiques, puis pour chaque tableau, choisi une valeur conservatrice parmi tous les effets du plan relatifs à ce tableau. Cette valeur choisie a ensuite été utilisée pour générer le tableau qui peut alors s'appliquer à l'ensemble complet des caractéristiques.

Les effets de plan, les tailles d'échantillon et les comptes de population qui ont servi à produire les tableaux de la variabilité d'échantillonnage approximative de même que les tableaux sont disponibles à l'annexe E. Tous les coefficients de variation sont *approximatifs* dans les tableaux de la variabilité d'échantillonnage approximative et ils ne doivent donc pas être considérés comme des valeurs exactes. Les possibilités concernant le calcul d'un coefficient de variation exact sont discutées dans la sous-section 11.7.

Rappel : Tel qu'indiqué dans les lignes directrices à la section 10.4, si le nombre d'observations sur lesquelles une estimation est basée est inférieur à 30, l'estimation pondérée ne doit pas être diffusée, quelle que soit la valeur de son coefficient de variation. Les coefficients de variation d'estimations basées sur des échantillons de petite taille sont trop imprévisibles pour être adéquatement représentés dans les tableaux.

11.1 Comment utiliser les tableaux de CV pour les estimations de type nominal

Les règles suivantes devraient permettre à l'utilisateur de calculer à partir des tableaux de la variabilité d'échantillonnage, les coefficients de variation approximatifs d'estimations relatives au nombre, à la proportion ou au pourcentage de personnes dans la population observée qui possèdent une caractéristique donnée ainsi que des rapports et des écarts entre ces estimations.

Règle 1 : Estimations du nombre de personnes possédant une caractéristique donnée (agrégats)

Le coefficient de variation dépend uniquement de la taille de l'estimation elle-même. Dans le tableau de variabilité d'échantillonnage correspondant à la région appropriée, il faut repérer l'estimation calculée dans la colonne d'extrême gauche (intitulée « Numérateur du pourcentage ») et suivre les astérisques (s'il y en a) de gauche à droite jusqu'au premier nombre. Puisque toutes les valeurs possibles de l'estimation ne sont pas disponibles dans cette colonne, il faut prendre la valeur la plus petite qui s'en rapproche le plus (par exemple, si l'estimation vaut 1 700 et que les deux valeurs disponibles dans la colonne « Total » se rapprochant le plus de 1 700 sont 1 000 et

2 000, il faut choisir 1 000) Ce nombre constitue le coefficient de variation approximatif pour l'estimation en question.

Règle 2 : Estimations de proportions ou de pourcentages de personnes possédant une caractéristique donnée

Le coefficient de variation d'une proportion (ou d'un pourcentage) estimée dépend à la fois de l'ordre de grandeur de cette proportion et de l'ordre de grandeur du numérateur utilisé dans le calcul de la proportion. Les proportions estimées sont relativement plus fiables que les estimations correspondantes du numérateur de la proportion lorsque celle-ci est fondée sur un sous-ensemble de la population. Cela est dû au fait que les coefficients de variation des estimations du dernier type sont basés sur le chiffre le plus élevé dans une rangée d'un tableau particulier, tandis que les coefficients de variation des estimations du premier type sont basés sur un chiffre quelconque de cette même rangée (pas nécessairement le plus élevé). (Il convient de noter que dans les tableaux, la valeur des coefficients de variation décroît de gauche à droite sur une même ligne.) Par exemple, la proportion estimative de personnes qui fument tous les jours parmi les fumeurs est plus fiable que le nombre estimatif de personnes qui fument tous les jours.

Lorsque la proportion (ou le pourcentage) est fondée sur la population totale de la région géographique à laquelle le tableau s'applique, le coefficient de variation de la proportion est égal à celui du numérateur de la proportion. Dans ce cas-ci, cela équivaut à appliquer la règle 1.

Lorsque la proportion (ou le pourcentage) est fondée sur un sous-ensemble de la population totale (p. ex., les personnes qui fument), il faut se reporter à la proportion (haut du tableau) et au numérateur de la proportion ou du pourcentage (côté gauche du tableau). Puisque toutes les valeurs possibles de la proportion et du numérateur ne sont pas disponibles, il faut, dans les deux cas, prendre la valeur la plus petite qui s'en rapproche le plus (par exemple, si la proportion est de 23 % et que les deux valeurs disponibles sur la ligne « Pourcentage estimé » s'en rapprochant le plus sont 20 % et 25 %, il faut choisir 20 %). Le coefficient de variation se trouve à l'intersection de la ligne et de la colonne appropriée. Si, à cet endroit, on retrouve des astérisques, il faut prendre le premier chiffre que l'on retrouve à la droite de ces astérisques.

Règle 3 : Estimations des différences entre des agrégats ou des pourcentages

L'erreur-type d'une différence entre deux estimations est à peu près égale à la racine carrée de la somme des carrés de chaque erreur-type considérée séparément. L'erreur-type d'une différence ($\hat{d} = \hat{X}_2 - \hat{X}_1$) est donc :

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

où \hat{X}_1 représente l'estimation 1, \hat{X}_2 l'estimation 2, et α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et \hat{X}_2 respectivement. Le coefficient de variation de \hat{d} est donné par $\sigma_{\hat{d}} / \hat{d}$. Cette formule donne un résultat exact pour ce qui est de la différence entre des sous-populations indépendantes. Dans le cas où les sous-populations ne sont pas indépendantes, ce calcul mènera à une

surestimation ou une sous-estimation selon que la relation (ou corrélation) entre ces deux sous-populations est positive ou négative.

Règle 4 : Estimations de rapports

Si le numérateur est un sous-ensemble du dénominateur, il faut convertir le rapport en pourcentage et appliquer la règle 2. Ce serait le cas, par exemple, si le dénominateur est le nombre de personnes qui fument et le numérateur est le nombre de personnes qui fument tous les jours parmi celles qui fument.

Si le numérateur n'est pas un sous-ensemble du dénominateur (par exemple, le rapport du nombre de personnes qui fument tous les jours ou à l'occasion au nombre de personnes qui ne fument pas du tout), l'écart-type du rapport entre les estimations est à peu près égal à la racine carrée de la somme des carrés de chaque coefficient de variation pris séparément multipliée par \hat{R} , où \hat{R} est le rapport des estimations ($\hat{R} = \hat{X}_1 / \hat{X}_2$). L'erreur-type d'un rapport est donc :

$$\sigma_{\hat{R}} = \hat{R} \sqrt{\alpha_1^2 + \alpha_2^2}$$

où α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et \hat{X}_2 respectivement.

Le coefficient de variation de \hat{R} est donné par $\sigma_{\hat{R}} / \hat{R} = \sqrt{\alpha_1^2 + \alpha_2^2}$. La formule tend à surestimer l'erreur si \hat{X}_1 et \hat{X}_2 sont corrélés positivement et à sous-estimer l'erreur si \hat{X}_1 et \hat{X}_2 sont corrélés négativement.

Règle 5 : Estimations des différences entre des rapports

Dans ce cas-ci, les règles 3 et 4 sont combinées. On commence par calculer les coefficients de variation des deux rapports au moyen de la règle 4, puis le coefficient de variation de leur différence au moyen de la règle 3.

11.2 Exemples d'utilisation des tableaux de CV pour des estimations de type nominal

Les exemples réels suivants ont pour but d'aider les utilisateurs à appliquer les règles décrites ci-dessus

Exemple 1 : Estimations du nombre de personnes possédant une caractéristique donnée (agrégats)

Supposons qu'un utilisateur estime à 4 722 617 le nombre de personnes qui fument tous les jours au Canada. Comment l'utilisateur fait-il pour déterminer le coefficient de variation de cette estimation?

- 1) Se reporter au tableau de CV pour le CANADA.
- 2) L'agrégat estimé (4 722 617) ne figure pas dans la colonne de gauche (la colonne «Numérateur du pourcentage»); il faut donc utiliser le nombre le plus petit qui s'en rapproche le plus, soit 4 000 000.
- 3) Le coefficient de variation d'un agrégat estimé (exprimé en pourcentage) est la première entrée sur cette ligne (à part les astérisques), soit 1,0 %.
- 4) Le coefficient de variation approximatif de l'estimation est donc 1,0 %. Par conséquent, selon les lignes directrices présentées à la section 10.4, l'estimation selon laquelle 4 722 617 personnes fument tous les jours peut être diffusée sans réserve.

Exemple 2 : Estimations de proportions ou de pourcentages de personnes possédant une caractéristique donnée

Supposons qu'un utilisateur estime à $4\,722\,617 / 6\,081\,453 = 77,7\%$ le pourcentage de personnes, parmi les fumeurs, qui fument tous les jours au Canada. Comment l'utilisateur fait-il pour déterminer le coefficient de variation de cette estimation?

- 1) Se reporter au tableau de CV pour le CANADA.
- 2) Parce que l'estimation est un pourcentage basé sur un sous-ensemble de la population totale (c.-à-d. les personnes qui fument tous les jours ou à l'occasion), il faut utiliser à la fois le pourcentage (77,7 %) et la partie numérateur du pourcentage (4 722 617) pour déterminer le coefficient de variation.
- 3) Le numérateur (4 722 617) ne figure pas dans la colonne de gauche (la colonne «Numérateur du pourcentage»); il faut donc utiliser le nombre le plus petit qui s'en rapproche le plus, soit 4 000 000. De même, l'estimation du pourcentage ne figure pas parmi les en-têtes de colonnes; il faut donc utiliser le nombre le plus petit qui s'en rapproche le plus, soit 70,0 %.
- 4) Le nombre qui se trouve à l'intersection de la ligne et de la colonne utilisées, soit 0,6 %, est le coefficient de variation (exprimé en pourcentage) à employer.
- 5) Le coefficient de variation de l'estimation est donc 0,6 %. Par conséquent, selon les lignes directrices présentées à la section 10.4, l'estimation selon laquelle 77,7 % des gens qui fument le font tous les jours peut être diffusée sans réserve.

Exemple 3 : Estimations des différences entre des agrégats ou des pourcentages

Supposons qu'un utilisateur estime que, parmi les hommes, $2\,535\,367/13\,078\,499 = 19,4\%$ fument tous les jours (estimation 1), alors que chez les femmes, ce pourcentage est estimé à $2\,187\,250/13\,476\,931 = 16,2\%$ (estimation 2). Comment l'utilisateur fait-il pour déterminer le coefficient de variation de la différence entre ces deux estimations?

- 1) À l'aide du tableau de CV pour le CANADA, utilisé de la même façon que dans l'exemple 2, vous établissez à 1,5 % le CV de l'estimation 1 (exprimé en pourcentage) et à 1,5 % le CV de l'estimation 2 (exprimé en pourcentage).
- 2) Selon la règle 3, l'erreur-type pour une différence ($\hat{d} = \hat{X}_2 - \hat{X}_1$) est :

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

où \hat{X}_1 est l'estimation 1, \hat{X}_2 est l'estimation 2, et α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et \hat{X}_2 respectivement. L'erreur-type de la différence $\hat{d} = (0,194 - 0,162) = 0,032$ est donc :

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(0,194)(0,015)]^2 + [(0,162)(0,015)]^2} \\ &= 0,004\end{aligned}$$

- 3) Le coefficient de variation de \hat{d} est donné par $\sigma_{\hat{d}} / \hat{d} = 0,004/0,032 = 0,125$.
- 4) Le coefficient de variation approximatif de la différence entre les estimations est donc 12,5 % (exprimé en pourcentage). Par conséquent, toujours selon les lignes directrices présentées à la section 10.4, cette estimation peut être publiée sans réserve.

Exemple 4 : Estimations de rapports

Supposons qu'un utilisateur estime à 4 722 617 le nombre de personnes qui fument tous les jours et à 1 358 836 le nombre de celles qui fument à l'occasion. L'utilisateur veut comparer ces deux estimations sous la forme d'un rapport. Comment fait-il pour déterminer le coefficient de variation de cette estimation?

- 1) Tout d'abord, cette estimation est une estimation de rapport, où le numérateur de l'estimation ($= \hat{X}_1$) est le nombre de personnes qui fument à l'occasion. Le dénominateur de l'estimation ($= \hat{X}_2$) est le nombre de personnes qui fument tous les jours.
- 2) Se reporter au tableau de CV pour le CANADA.

- 3) Le numérateur de cette estimation de rapport est 1 358 836. Le nombre le plus petit qui se rapproche le plus de ce nombre est 1 000 000. Le coefficient de variation de cette estimation (exprimé en pourcentage) est la première entrée sur cette ligne (à part les astérisques), soit 2,3 %.
- 4) Le dénominateur de cette estimation de rapport 4 722 617. Le nombre le plus petit qui se rapproche le plus de ce nombre est 4 000 000. Le coefficient de variation de cette estimation (exprimé en pourcentage) est la première entrée sur cette ligne (à part les astérisques), soit 1,0 %.
- 5) Le coefficient de variation approximatif de l'estimation du rapport est donc donné par la règle 4,

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^2 + \alpha_2^2},$$

c'est-à-dire,

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(0,023)^2 + (0,01)^2} \\ &= 0,025\end{aligned}$$

où α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et \hat{X}_2 respectivement. Le rapport des personnes qui fument occasionnellement à celles qui fument tous les jours est 1 358 836/4 722 617, soit 0,29:1. Le coefficient de variation de cette estimation est 2,5 % (exprimé en pourcentage); selon les lignes directrices présentées à la section 10.4, l'estimation peut donc être diffusée sans réserve.

11.3 Comment utiliser les tableaux de CV pour calculer les limites de confiance

Bien que les coefficients de variation soient largement utilisés, l'intervalle de confiance d'une estimation représente une mesure plus intuitive de l'erreur d'échantillonnage. Un intervalle de confiance est une façon d'énoncer la probabilité que la valeur vraie de la population se situe dans une plage de valeurs données. Par exemple, un intervalle de confiance de 95 % peut être décrit comme suit : si l'échantillonnage de la population se répète à l'infini, chacun des échantillons donnant un nouvel intervalle de confiance pour une estimation, l'intervalle contiendra la valeur vraie de la population dans 95 % des cas.

Une fois déterminée l'erreur-type d'une estimation, on peut calculer des intervalles de confiance pour les estimations en partant de l'hypothèse qu'en procédant à un échantillonnage répété de la population, les diverses estimations obtenues pour une caractéristique de la population sont réparties selon une distribution normale autour de la valeur vraie de la population. Selon cette hypothèse, il y a environ 68 chances sur 100 que l'écart entre une estimation de l'échantillon et la valeur vraie de la population soit inférieur à une erreur-type, environ 95 chances sur 100 que l'écart soit inférieur à deux erreurs-types et environ 99 chances sur 100 que l'écart soit inférieur à trois erreurs-types. On appelle ces différents degrés de confiance des niveaux de confiance.

L'intervalle de confiance d'une estimation \hat{X} est généralement exprimé sous la forme de deux nombres, l'un étant inférieur à l'estimation et l'autre supérieur à celle-ci, sous la forme ($\hat{X} - k$, $\hat{X} + k$), où k varie selon le niveau de confiance désiré et l'erreur d'échantillonnage de l'estimation.

On peut calculer directement les intervalles de confiance d'une estimation à partir des tableaux de la variabilité d'échantillonnage approximative, en trouvant d'abord dans le tableau approprié le coefficient de variation de l'estimation \hat{X} , puis en utilisant la formule suivante pour obtenir l'intervalle de confiance CI correspondant :

$$CI_X = [\hat{X} - z \hat{X} \alpha_{\hat{X}}, \hat{X} + z \hat{X} \alpha_{\hat{X}}]$$

où $\alpha_{\hat{X}}$ est le coefficient de variation trouvé pour \hat{X} , et

$z = 1$ si l'on désire un intervalle de confiance de 68 %
 $z = 1,6$ si l'on désire un intervalle de confiance de 90 %
 $z = 2$ si l'on désire un intervalle de confiance de 95 %
 $z = 3$ si l'on désire un intervalle de confiance de 99 %

Note : Les lignes directrices concernant la diffusion des estimations de la section 10.4 s'appliquent aussi aux intervalles de confiance. Par conséquent, si l'estimation ne peut être diffusée, alors l'intervalle de confiance ne peut l'être lui non plus.

11.4 Exemple d'utilisation de tableaux de CV pour obtenir des limites de confiance

Voici la marche à suivre pour calculer un intervalle de confiance de 95 % pour la proportion estimée de personnes qui fument tous les jours parmi celles qui fument (d'après l'exemple 2 de la sous-section 11.2).

$$\hat{X} = 0,777$$

$$z = 2$$

$$\alpha_{\hat{X}} = 0,006 \text{ est le coefficient de variation de cette estimation selon les tableaux.}$$

$$CIX = \{0,777 - (2) (0,777) (0,006), 0,777 + (2) (0,777) (0,006)\}$$

$$CIX = \{0,768, 0,786\}$$

11.5 Comment utiliser les tableaux de CV pour effectuer un test Z

On peut aussi utiliser les erreurs-types pour effectuer des tests d'hypothèses, une technique qui permet de faire la distinction entre les paramètres d'une population à l'aide d'estimations basées sur un échantillon. Ces estimations peuvent être des nombres, des moyennes, des pourcentages, des rapports, etc. Les tests peuvent être effectués à divers niveaux de signification; un niveau de signification est la probabilité de conclure que les caractéristiques sont différentes quand, en fait, elles sont identiques.

Supposons que \hat{X}_1 et \hat{X}_2 sont des estimations basées sur un échantillon pour deux caractéristiques voulues. Supposons aussi que l'erreur-type de la différence $\hat{X}_1 - \hat{X}_2$ est $\sigma_{\hat{d}}$. Si $z = (\hat{X}_1 - \hat{X}_2) / \sigma_{\hat{d}}$ est compris entre -2 et 2, alors on ne peut tirer aucune conclusion à propos de la différence entre les caractéristiques au niveau de signification de 5 %. Toutefois, si ce rapport est inférieur à -2 ou supérieur à +2, la différence observée est significative au niveau de 0,05.

11.6 Exemple d'utilisation des tableaux de CV pour effectuer un test Z

Supposons que nous voulons tester, au niveau de signification de 5 %, l'hypothèse selon laquelle il n'y a pas de différence entre la proportion d'hommes qui fument tous les jours et cette même proportion chez les femmes. Dans l'exemple 3 de la sous-section 11.2, nous avons déterminé que l'erreur-type de la différence entre ces deux estimations est égale à 0,004. Par conséquent,

$$z = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{0,194 - 0,162}{0,004} = \frac{0,032}{0,004} = 8$$

Puisque $z = 8$ est supérieur à 2, on doit conclure qu'il existe une différence significative entre les deux estimations au niveau de signification de 0,05. À noter que les deux sous-groupes comparés sont considérés comme étant indépendants, ce qui fait en sorte que le résultat du test est valide.

11.7 Variances ou coefficients de variation exacts

Tous les coefficients de variation qui figurent dans les tableaux de la variabilité d'échantillonnage approximative (tableaux de CV) sont effectivement approximatifs, donc, non officiels.

Le calcul de variance ou coefficient de variation exact n'est pas chose évidente puisqu'il n'existe pas de formule mathématique simple pouvant prendre en compte de tous les aspects du plan d'échantillonnage et de la pondération de l'ESCC. On doit donc avoir recours à d'autres méthodes pour estimer ces mesures de précisions, telles que des méthodes par rééchantillonnage. Parmi celles-ci, la méthode du bootstrap est celle recommandée pour l'analyse des données de l'ESCC.

Le calcul de coefficients de variation (ou tout autre mesure de précision) fait à l'aide de la méthode du bootstrap nécessite toutefois l'accès à de l'information considérée confidentielle qui n'est évidemment pas disponible dans le fichier de microdonnées à grande diffusion. Le calcul doit donc se faire à l'aide du fichier maître. L'accès au fichier maître est discuté à la section 12.3.

Pour le calcul de coefficients de variation, il est conseillé d'utiliser la méthode du bootstrap. Un programme macro, appelé le "Bootvar", a été développé pour faciliter le calcul à l'aide de la méthode bootstrap. Le programme Bootvar est offert en formats SAS et SPSS, et est constitué de macros qui calculent les variances de totaux, ratios, différences entre ratios, et pour des régressions linéaires et logistiques.

Les raisons pour lesquelles un utilisateur pourrait souhaiter connaître la précision exacte de ses estimations sont diverses. En voici quelques-unes.

Premièrement, si un utilisateur désire obtenir des estimations à un niveau géographique autre que ceux présentés dans les tableaux (par exemple, au niveau urbain ou rural), l'utilisation des tableaux de CV publiés ne convient pas parfaitement. Néanmoins, on peut obtenir les coefficients de variation de ce type d'estimations en appliquant la méthode d'estimation par domaine, au moyen du programme de calcul de la variance exacte (le "Bootvar").

Deuxièmement, si un utilisateur demande des analyses plus complexes, telles que des estimations de paramètres de modèles de régression linéaire ou logistique, les tableaux de CV ne pourront pas fournir les coefficients de variation pour ceux-ci. Certains progiciels statistiques courants permettent d'incorporer les poids d'échantillonnage aux analyses, mais, souvent, les variances produites ne tiennent pas bien compte de la stratification et de la mise en grappe de l'échantillon, contrairement à celles obtenues grâce au programme de calcul de la variance exacte.

Troisièmement, dans le cas de l'estimation de variables quantitatives, il est nécessaire d'utiliser des tableaux distincts pour déterminer l'erreur d'échantillonnage. Or, la plupart des variables de l'ESCC étant de type nominal, de tels tableaux n'ont pas été produits. Les utilisateurs qui souhaitent connaître les coefficients de variation de variables quantitatives peuvent néanmoins obtenir ces derniers grâce au programme de calcul de la variance réelle. À noter, toutefois, que le coefficient de variation d'un total quantitatif est généralement plus grand que celui de l'estimation de type nominal correspondante (c'est-à-dire, l'estimation du nombre de personnes qui contribuent à l'estimation quantitative). Si l'estimation de type nominal correspondante ne peut être diffusée, il en sera de même pour l'estimation quantitative. Par exemple, le coefficient de variation de l'estimation du nombre total de cigarettes que fument chaque jour les personnes qui fument tous les jours serait supérieur à celui de l'estimation correspondante du nombre de personnes qui fument tous les jours. Par conséquent, si on ne peut diffuser le coefficient de variation de cette dernière estimation, on ne pourra non plus diffuser celui de l'estimation quantitative correspondante.

Enfin, un utilisateur qui peut se servir des tableaux de CV, mais obtient ainsi un coefficient de variation compris dans la fourchette marginale (de 16,6 % à 33,3 %), devrait diffuser les estimations associées en y joignant une mise en garde aux utilisateurs quant à la variabilité d'échantillonnage élevée liée aux estimations. Dans ce cas, il serait bon de recalculer le coefficient de corrélation à l'aide du programme de variance exacte pour vérifier si ces estimations peuvent être diffusées sans mise en garde. Cette situation tient au fait que l'estimation des coefficients de variation grâce aux tableaux de la variabilité d'échantillonnage approximative est basée sur une

vaste gamme de variables et, donc, jugée grossière, alors que le programme de calcul de la variance réelle produit le coefficient de variation précis associé à la variable en question.

11.8 Seuils pour la diffusion des estimations relatives à l'ESCC

L'annexe E présente les tableaux indiquant les seuils de diffusion des totaux selon les estimations pour le Canada, les provinces, les régions sociosanitaires, les régions de CLSC ainsi que pour les différents groupes d'âges (pour le Canada seulement). Les estimations inférieures à la valeur indiquée dans la colonne «Marginal» ne peuvent en aucun cas être diffusées.

12. Utilisation du fichier

La présente section débute par une description de la variable de pondération présente sur le fichier de microdonnées à grande diffusion, et fournit des explications sur la façon de l'utiliser lorsqu'on effectue des totalisations. Suit une explication de la convention appliquée pour nommer les variables de l'ESCC. Ensuite, vient la description des diverses méthodes d'accès aux données que peuvent adopter les analystes.

12.1 Utilisation de la variable de pondération

La variable de pondération **WTSC_M** représente le poids d'échantillonnage pour le cycle 2.1 de l'ESCC. Pour un répondant donné, ce poids d'échantillonnage peut être interprété comme étant le nombre de personnes que le répondant représente dans la population. Ce poids doit être utilisé en tout temps dans les calculs d'estimations statistiques, afin de permettre l'inférence à l'échelle de la population. La production de résultats non pondérés n'est pas recommandée. La répartition de l'échantillon, de même que les détails du plan de sondage, peuvent entraîner des résultats biaisés qui ne représentent pas correctement la population. Pour une description plus détaillée du calcul de ce poids, consulter la section 8 sur la pondération.

12.2 Convention appliquée pour nommer les variables

On a adopté, pour nommer les variables de l'ESCC, une convention 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 1.1, 1.2...) dans le nom, et permettre de repérer facilement les variables conceptuellement identiques d'un cycle à l'autre 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.

12.2.1 Structure élémentaire des noms des variables de l' ESCC

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

Positions 1 à 3 :	Nom du module/de la section du questionnaire
Position 4 :	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 202, module Usage du tabac, cycle 2.1, c'est-à-dire SMKC_202 est la suivante :

Positions 1 à 3 : SMK Module de la dépression
Position 4 : C Cycle 2.1
Position 5 : _ (_ = données recueillies)
Position 6 à 8 : 202 numéro de la question et option de réponse

12.2.2 Positions 1 à 3 : Nom de la variable/section du questionnaire

On se sert des valeurs suivantes pour la composante du nom de la variable correspondant à la section du questionnaire :

ADM	Administration	LEI	Activités de loisir
ALC	Consommation d'alcool	MAM	Mammographie
ALD	Dépendance à l'égard de l'alcool	MAS	Contrôle
BPC	Tension artérielle	MED	Consommation de médicaments
BRX	Examen des seins	MEX	Expérience de la maternité
BSX	Auto-examen des seins	NDE	Dépendance à la nicotine
CCC	Problèmes de santé chroniques	OH1	Santé bucco-dentaire - commun
CCS	Dépistage du cancer du côlon et du rectum	OH2	Santé bucco-dentaire - optionnel
CIH	Changements pour améliorer la santé	ORG	Organismes bénévoles
CMH	Consultations des spécialistes de la santé mentale	PAC	Activités physiques
CPG	Jeu pathologique	PAP	Test Papanicolaou
DEN	Visites chez le dentiste	PAS	Satisfaction des patients
DHH	Données démographiques et composition du ménage	PCU	Examen général
DIQ	Détresse et état de santé mentale (Québec)	PSA	Test de l'antigène spécifique prostatique
DIS	Détresse	QMD	Consommation de médicaments (Québec)
DPS	Dépression	RAC	Limitation des activités
DRV	Conduite automobile et sécurité	REP	Mouvement répétitif
DSU	Utilisation de compléments vitaminiques	SAC	Activités sédentaires
EDU	Niveau de scolarité	SAM	Identificateurs d'échantillon
ETA	Troubles de l'alimentation	SCA	Outils pour arrêter de fumer
ETS	Exposition à la fumée des autres	SCH	Étapes du changement (usage du tabac)
FDC	Choix alimentaires	SDC	Renseignements sociodémographiques
FIN	Insécurité alimentaire	SFE	Estime de soi
FLU	Vaccination contre la grippe	SFR	État de santé - SF-36

FVC	Consommation de fruits et de légumes	SMK	Usage du tabac
GEN	État de santé général	SPC	Consultation d'un médecin (usage du tabac)
GEO	Identificateurs géographiques	SSM	Soutien social
HCS	Satisfaction à l'égard du système de soins de santé	SUI	Pensées suicidaires et tentatives de suicide
HCU	Utilisation des soins de santé	SWA	Satisfaction concernant la disponibilité des services de santé
HMC	Soins à domicile	SWL	Satisfaction à l'égard de la vie
HMS	Sécurité à la maison	SXB	Comportement sexuel
HUI	Indice de l'état de santé (HUI)	TAL	Variantes du tabagisme
HWT	Taille et poids	TWD	Incapacité au cours des deux dernières semaines
IDG	Drogues illicites	UPE	Utilisation de protections
INC	Revenu	WST	Stress au travail
INJ	Blessures	WTS	Poids de sondage
INS	Couverture d'assurance	YSM	Usage du tabac chez les jeunes
LBF	Population active		

12.2.3 Position 4 : Cycle

Cycle Description

- A** Cycle 1.1 : Enquête sur la santé dans les collectivités canadiennes
:enquête à l'échelle régionale, échantillon stratifié selon la région sociosanitaire
:contenu commun et contenu optionnel sélectionnés par les régions sociosanitaires
:estimations à l'échelle régionale (régions sociosanitaires), (provincial, territorial et national)
- B** Cycle 1.2 : Enquête sur la santé dans les collectivités canadiennes, santé mentale et bien-être
:enquête à l'échelle provinciale
:contenu thématique et contenu général supplémentaire
:estimations aux échelle provinciale, territoriale et nationale
- C** Cycle 2.1 : Enquête sur la santé dans les collectivités canadiennes
:enquête à l'échelle régionale, échantillon stratifié selon la région socio-sanitaire
:contenu commun et contenu optionnel sélectionnés par les régions sociosanitaires
:estimations à l'échelle régionale (régions sociosanitaires), (provincial, territorial et national)

12.2.4 Position 5 : Type de variable

–	Variable collectée	Variable qui figure directement sur le questionnaire
C	Variable codée	Variable codée à partir d'une ou de plusieurs variables collectées (par exemple, code de la Classification type des industries (CTI))
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)

12.2.5 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 au moyen d'un groupe de deux chiffres. Par exemple, la question Q01A du questionnaire devient simplement 01A et la question Q15, simplement 15.

Parfois, certaines questions comportent plusieurs réponses alors la position finale dans la séquence du nom de la variable est représentée par une lettre. Pour ce genre de questions, de nouvelles variables sont créées dans le but de différencier un “oui” d'un “non” pour chaque possibilité de réponse. Par exemple, si la question Q2 a 4 réponses possibles, les nouvelles questions seraient Q2A pour la première possibilité, Q2B pour la deuxième, Q2C pour la troisième et ainsi de suite. Si seulement les options 2 et 3 sont choisies, alors Q2A = Non, Q2B = Oui, Q2C = Oui et Q2D = Non.

12.3 Accès au fichier maître

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 sévères de sécurité et de confidentialité, conformément à la *Loi sur la statistique*. Pour s'assurer du respect de ces normes, chaque fichier de microdonnées est soumis 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. Fréquemment, ces variables sont les plus essentielles à l'analyse complète des données d'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 afin de bien rentabiliser l'investissement statistique.

Une première méthode offerte à tous les utilisateurs consiste à demander au personnel des Services personnalisés à la clientèle de la Division de la statistique de la santé de produire des totalisations personnalisées. Ce service permet aux utilisateurs qui ne savent pas se servir de logiciels de totalisation d'obtenir des résultats personnalisés. Les résultats sont filtrés pour s'assurer qu'ils sont conformes aux normes de confidentialité et de fiabilité avant d'être diffusés. Ce service est offert contre remboursement des frais.

Deuxièmement, le Programme des centres de données de recherche permet aux chercheurs de soumettre à Statistique Canada un projet de recherche fondé sur les données des fichiers maîtres. Un ensemble particulier de règles est appliqué afin de décider quels projets seront acceptés. Lorsque le projet est accepté, le chercheur est considéré comme étant "réputé employé" par Statistique Canada pour la durée de l'étude et se voit accorder l'accès au fichier maître de l'enquête dans des locaux désignés de Statistique Canada. Pour plus de renseignements, consultez la page web suivante : http://www.statcan.ca/francais/rdc/index_f.htm

En dernier lieu, le service de télé-accès aux fichiers maîtres de l'enquête est un moyen d'accéder à ces données si il est impossible de passer par un Centre. On peut fournir à l'acheteur d'un produit de microdonnées un fichier maître "fictif" d'essai et le cliché d'enregistrement correspondant. Grâce à ces outils, il peut mettre au point son propre ensemble de programmes analytiques en se servant du fichier fictif pour confirmer que les routines fonctionnent convenablement. Il ne lui reste plus qu'à envoyer le code pour les totalisations personnalisées par courrier électronique à cchs-escc@statcan.ca. Le code est transmis au réseau interne protégé de Statistique Canada et traité en regard du fichier maître approprié de données du cycle 2.1 de l'ESCC. Les estimations générées seront communiquées à l'utilisateur, sujet aux directives sur l'analyse et la communication des données tel qu'exposé dans les grandes lignes à la section 10 de ce document. Les résultats sont filtrés pour vérifier s'ils sont conformes aux normes de confidentialité et de fiabilité, puis, les données de sortie sont renvoyées au client. Ce service est gratuit.