Guide de l’utilisateur de microdonnées

pour le fichier de microdonnées à grande diffusion

Enquête canadienne sur les problèmes juridiques (ECPJ)

2021

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements
Toute demande de renseignements concernant le présent produit et les statistiques ou services connexes doit être adressée à :

Statistique Canada
Service à la clientèle
Centre de l'intégration et du développement des données sociales
Téléphone : 613-951-3321 ou numéro sans frais : 1-800-461-9050
Courriel : csdid-info-cidds@statcan.gc.ca

Accéder aux microdonnées et passer une commande
L'Enquête canadienne sur les problèmes juridiques (ECPJ) de 2021 permet de produire deux types de fichiers de microdonnées : un fichier maître et un fichier de microdonnées à grande diffusion (FMGD).

Fichier maître
Le fichier maître contient toutes les variables et tous les enregistrements dans le champ de l'enquête recueillis au cours d'une période de collecte. Ce fichier est accessible à Statistique Canada pour un usage interne et dans les centres de données de recherche (CDR) de Statistique Canada. Il peut aussi faire l'objet de demandes de totalisations personnalisées.

Centres de données de recherche
Le programme des CDR permet aux chercheurs d'utiliser les données d'enquête contenues dans les fichiers maîtres dans un environnement sécurisé, lequel est situé au sein d'une université. Plusieurs universités disposent de CDR, et ce, partout au pays. Les chercheurs doivent soumettre des propositions de recherche qui, si elles sont acceptées, leur permettront d'avoir accès aux CDR. Pour plus de renseignements, consultez la page Web suivante : https://www.statcan.gc.ca/fr/microdonnees/centres-donnees.

Totalisations personnalisées
Une autre méthode d'accès au fichier maître consiste à offrir à tous les utilisateurs de faire appel au personnel du Service à la clientèle du Centre de l'intégration et du développement des données sociales de Statistique Canada pour produire des totalisations personnalisées. Ce service offert moyennant le recouvrement des coûts permet aux utilisateurs qui n'ont pas les connaissances nécessaires à l'utilisation 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. Pour obtenir plus de renseignements, veuillez communiquer avec le Service à la clientèle.

Fichier de microdonnées à grande diffusion
Le fichier de microdonnées à grande diffusion (FMGD) est créé à partir du fichier maître au moyen d'une technique établissant un équilibre entre le besoin d'assurer la confidentialité des répondants et le besoin de produire les données les plus utiles possible. Le FMGD doit respecter des normes de sécurité et de confidentialité rigoureuses exigées par la *Loi sur la statistique* avant d'être rendu public. Pour veiller au respect de ces normes, chaque FMGD est soumis à un processus formel d'examen et d'approbation par un comité de direction de Statistique Canada. Les variables les plus susceptibles de permettre l'identification d'une personne sont supprimées du fichier de données ou agrégées pour former de plus vastes catégories.

Pour obtenir une copie du FMGD, veuillez contacter Infostats ou consulter le site Web de Statistique Canada.

Initiative de démocratisation des données
L'Initiative de démocratisation des données (IDD) permet aux étudiants et aux chercheurs d'utiliser les FMGD dans plusieurs universités partout au Canada. Pour obtenir plus de renseignements, veuillez consulter la page Web suivante : https://www.statcan.gc.ca/fr/microdonnees/idd.

Table des matières

1.0 Description de l'enquête

L'Enquête canadienne sur les problèmes juridiques (ECPJ) a été menée par Statistique Canada en 2021 pour le compte de Justice Canada. Cette enquête vise à déterminer les types de problèmes graves auxquels les personnes doivent faire face, la façon dont elles tentent de les résoudre et les répercussions de ces expériences sur leur vie. Les renseignements recueillis dans le cadre de l'enquête permettront de mieux comprendre les diverses mesures que les personnes prennent pour résoudre leurs problèmes, que ce soit par la voie de systèmes officiels comme les cours ou les tribunaux, ou encore par des voies informelles comme les stratégies d'autoassistance.

Le présent guide de l'utilisateur a été élaboré pour faciliter la manipulation du fichier de microdonnées à grande diffusion (FMGD) de l'enquête.

Les questions concernant l'ensemble de données ou son utilisation doivent être adressées à :

Statistique Canada
Service à la clientèle
Centre de l'intégration et du développement des données sociales
Téléphone : 613-951-3321 ou numéro sans frais : 1-800-461-9050
Courriel : csdid-info-cidds@statcan.gc.ca

Justice Canada
Division de la recherche et de la statistique
284, rue Wellington, 6e étage, Ottawa, ON K1A 0H8
Courriel : rsd.drs@justice.gc.ca

2.0 Concepts et définitions

2.1 Concepts et définitions de l'ECPJ

Les concepts et définitions utilisés dans le cadre de l'Enquête canadienne sur les problèmes juridiques (ECPJ) sont pour la plupart d'usage commun. Par contre, les termes employés dans le questionnaire de l'enquête qui figurent ci-dessous pourraient avoir des définitions différentes de celles utilisées dans d'autres contextes.

Le **harcèlement** se définit comme tout comportement inopportun et injurieux d'une personne envers une autre et dont l'auteur savait ou aurait raisonnablement dû savoir qu'un tel comportement pouvait offenser ou causer un préjudice.

La **discrimination** se définit comme le fait de réserver à quelqu'un un traitement différent ou inéquitable en raison d'une caractéristique personnelle ou d'une distinction, intentionnelle ou non, qui a pour effet d'imposer des désavantages non imposés à d'autres, ou d'empêcher ou de restreindre l'accès aux avantages offerts à d'autres membres de la société.

2.2 Élaboration du contenu

Le contenu du questionnaire a été élaboré en collaboration étroite avec Justice Canada. Les analystes du Centre canadien de la statistique juridique et de la sécurité des collectivités (CCSJSC) ont participé à l'élaboration du questionnaire de l'ECPJ, puisqu'ils sont les experts en matière de justice à Statistique Canada.

Les questionnaires d'enquêtes antérieures menées par Justice Canada ou le Forum canadien sur la justice civile ont été consultés dans la phase initiale d'élaboration du questionnaire de l'ECPJ. Toutefois, les résultats de l'ECPJ ne devraient pas être comparés à ceux des enquêtes antérieures, en partie en raison des différences notables entre les questionnaires.

Des partenaires internes et externes désignés par Justice Canada ont également été consultés tout au long de l'élaboration du contenu de l'enquête pour veiller à ce que les questions de l'enquête répondent autant que possible à leurs besoins en données.

Au cours de son élaboration, le questionnaire de l'enquête a été mis à l'essai à trois reprises par les experts du Centre de ressources en conception de questionnaires (CRCQ) de Statistique Canada. En 2019, deux rondes d'interviews cognitives ont été réalisées pour évaluer la faisabilité du contenu. La première s'est tenue en février et en mars 2019. On a alors mené 35 interviews cognitives individuelles, qui ont eu lieu dans quatre villes : Halifax, Winnipeg, Ottawa et Montréal. Les personnes interviewées comprenaient des francophones et des anglophones, des personnes ayant connu divers types de problèmes juridiques au cours des trois années précédentes et, dans la mesure du possible, des personnes ayant divers âges, genres, niveaux de scolarité, situations d'emploi et revenus.

La deuxième ronde d'essais qualitatifs s'est tenue à Ottawa en septembre 2019, où un total de 19 interviews ont été menées dans les deux langues officielles. Dans le cadre de ces interviews, les participants ont été invités à répondre au questionnaire, y compris à de nouvelles questions, en personne (comme dans le cas d'une interview téléphonique).

En se fondant sur les résultats des essais qualitatifs, on a apporté des changements au questionnaire, et le contenu a été mis au point de concert avec Justice Canada. Une version du questionnaire en format PDF a été conçue en collaboration avec des partenaires internes de Statistique Canada afin de reproduire les écrans de l'application électronique. La version en format PDF du questionnaire, et dans certains cas une section du questionnaire électronique, a fait l'objet d'une troisième ronde d'essais qualitatifs, laquelle a été menée par le CRCQ en mars 2020. Ces essais se sont tenus à Montréal, Ottawa et Whitehorse; un total de 12 interviews ont été menées en français et 15 interviews ont été menées en anglais.

Tous les changements requis à la suite des essais qualitatifs ont été apportés à la conception des écrans de la version PDF et communiqués à Justice Canada. Ensuite, une application pour le questionnaire électronique a été mise au point et rigoureusement mise à l'essai par Statistique Canada.

L'enquête est constituée des modules suivants :

- Renseignements sociodémographiques (première partie);
- Identification des problèmes;
- Pandémie de COVID-19 (Non disponible sur le FMGD);
- Questions propres aux problèmes connus au cours des trois années précédentes, comprend les modules suivants : Achats et services aux consommateurs, Employeur et travail, Dette ou argent qui vous est dû, Interaction avec la police, Famille et relations, Garde des enfants et responsabilités parentales, Harcèlement, Discrimination;
- Liens entre les problèmes (1 seule variable présentée sur le FMGD);
- Problème le plus grave;
- Aide pour faire face au problème le plus grave;
- Comprendre la portée du problème le plus grave;
- État actuel du problème le plus grave;
- Assistance légale pour faire face au problème le plus grave;
- Coûts associés au problème le plus grave;
- Répercussions socioéconomiques;

- Santé et problèmes sociaux;
- Renseignements sociodémographiques (deuxième partie).

Pour obtenir de plus amples renseignements, veuillez consulter le questionnaire de l'ECPJ.

3.0 Méthodologie de l'enquête

3.1 Population cible et population observée

La population cible de l'enquête comprend les personnes de 18 ans et plus résidant dans l'une des 10 provinces canadiennes, à l'exception des personnes vivant en établissement, dans un logement collectif ou dans une réserve indienne.

Un échantillon de 42 400 personnes a été sélectionné au hasard à partir de la base de sondage. Cet échantillon est constitué d'un échantillon représentatif de 29 972 personnes de la population générale et d'un suréchantillon de 12 428 Autochtones.

3.2 Plan d'échantillonnage

La base de sondage de l'Enquête canadienne sur les problèmes juridiques (ECPJ) était une liste de personnes, établie à partir du questionnaire détaillé du Recensement de 2016 et d'autres fichiers administratifs. L'échantillon était composé de personnes sélectionnées au hasard à partir de cette base de sondage. Ces personnes ont été invitées à participer à l'enquête.

La base de sondage de l'ECPJ a été stratifiée par province et selon le statut autochtone ou non autochtone. Les strates d'Autochtones ont été sous-stratifiées selon l'identité autochtone (Premières Nations, Métis et Inuits) afin d'améliorer la qualité des estimations selon l'identité autochtone.

La base de sondage contenait jusqu'à cinq numéros de téléphone (ligne fixe ou cellulaire) pour permettre d'effectuer des suivis téléphoniques auprès des répondants.

3.3 Taille de l'échantillon

L'échantillon initial était composé d'un échantillon principal de 30 000 unités de la population non autochtone et d'un suréchantillon de 12 400 unités pour la population autochtone. L'échantillon principal et le suréchantillon ont été stratifiés par province au moyen d'une répartition de Kish. Cependant, en raison du nombre relativement faible d'Inuits et de résidents de l'Île-du-Prince-Édouard, certaines strates ont été combinées, ce qui a mené à la répartition finale : 29 972 unités de la population non autochtone et 12 428 unités pour la population autochtone.

4.0 Instrument de collecte et collecte des données

Dans le cadre de l'Enquête canadienne sur les problèmes juridiques (ECPJ), les données ont été recueillies à l'aide d'un questionnaire électronique en ligne, avec suivi téléphonique pour les cas de non-réponse. La collecte des données au moyen du questionnaire électronique a débuté le 1er février 2021 et s'est terminée le 20 août 2021.

Une lettre d'invitation accompagnée d'une brochure au sujet de l'enquête a été envoyée par la poste aux personnes sélectionnées afin de les inviter à participer à l'enquête. Chaque lettre contenait un code d'accès sécurisé permettant au répondant d'accéder au questionnaire électronique et de le remplir en ligne.

Dans la lettre d'invitation, la brochure et les lettres de rappel, les participants étaient informés de la nature volontaire de l'enquête ainsi que du fait que leurs renseignements demeureraient strictement confidentiels.

Quatre lettres de rappel, y compris une liste de ressources, ont été envoyées dans le but d'accroître le taux de réponse.

À partir du 15 mars 2021, des intervieweurs formés ont mené des interviews téléphoniques assistées par ordinateur (ITAO) à partir des bureaux régionaux pour tenter de contacter les personnes sélectionnées et les inciter à répondre à l'enquête. Au besoin, les intervieweurs ont pris en note tout renseignement ou autre numéro de téléphone permettant de joindre la personne sélectionnée. Les interviews par personne interposée n'étaient pas permises en raison du sujet délicat de l'enquête.

5.0 Traitement des données

Le traitement permet de transformer les réponses obtenues au cours de la collecte en un format qui se prête à la totalisation et à l'analyse des données. Le traitement comprend toutes les activités de manipulation des données, qu'elles soient automatisées ou manuelles, après la collecte et avant l'estimation.

5.1 Saisie des données

5.1.1 Questionnaire électronique

Pour ce qui est du questionnaire électronique, les réponses aux questions de l'enquête ont été saisies directement par le répondant. L'emploi de questionnaires électroniques réduit les délais et les coûts de traitement associés à la saisie des données, aux erreurs de transcription et à la transmission des données. Les réponses ont ensuite été transmises de façon sécuritaire au bureau central de Statistique Canada à Ottawa, au moyen de protocoles de chiffrement, de pare-feux et de couches de chiffrement conformes aux normes de l'industrie.

Une vérification des données a été effectuée directement au moment où le questionnaire électronique a été rempli. Lorsque les renseignements saisis étaient hors limites des valeurs attendues (trop faibles ou trop élevés), ou qu'ils entraient en contradiction avec des renseignements saisis auparavant, le répondant était invité à vérifier les renseignements fournis au moyen de messages s'affichant à l'écran de l'ordinateur. Toutefois, les répondants avaient la possibilité de passer outre aux vérifications et de sauter des questions s'ils ne connaissaient pas les réponses ou refusaient de répondre. Par conséquent, les données ont été soumises à d'autres processus de vérification après avoir été transmises au bureau central. Enfin, lorsque les données électroniques ont été reçues, elles ont été converties en fichiers texte lisibles.

5.1.2 Interviews téléphoniques assistées par ordinateur

En ce qui concerne les interviews téléphoniques assistées par ordinateur (ITAO), les réponses aux questions de l'enquête ont été saisies par l'intervieweur lors de l'interview au moyen d'un questionnaire électronique. L'emploi de questionnaires électroniques réduit les délais de traitement et les coûts associés à la saisie des données, aux erreurs de transcription et à la transmission des données. Les réponses fournies par les répondants ont ensuite été transmises de façon sécuritaire à Statistique Canada à Ottawa, au moyen de protocoles de chiffrement, de pare-feux et de couches de chiffrement conformes aux normes de l'industrie.

Une vérification des données a été effectuée directement au moment de l'interview. Lorsque les renseignements saisis sont hors limites des valeurs attendues (trop faibles ou trop élevés), ou qu'ils entrent en contradiction avec des renseignements saisis auparavant, l'intervieweur voit apparaître à l'écran de l'ordinateur des messages lui demandant de clarifier les renseignements auprès du répondant. Cependant, pour certaines questions, l'intervieweur a la possibilité de passer outre aux contrôles et de sauter des questions si le répondant ne connaît pas la réponse ou refuse de répondre. Par conséquent, les données des réponses ont été soumises à d'autres processus de vérification après avoir été transmises au bureau central.

5.2 Vérifications

Les vérifications peuvent être effectuées à plusieurs étapes du processus d'enquête, allant de simples vérifications préliminaires effectuées par l'application de collecte à des vérifications automatisées plus complexes effectuées à l'étape du traitement, après la saisie de données. Les règles de vérification sont généralement déterminées par ce qui peut être logique ou valide, compte tenu :

- des connaissances du spécialiste du domaine;
- d'autres enquêtes ou données connexes;
- de la structure du questionnaire et de ses questions;de la théorie statistique.

Il existe trois principales catégories de vérification : les vérifications de la validité, de la cohérence et de la répartition. Les vérifications de la validité s'attardent à la syntaxe des réponses et permettent entre autres de vérifier que les données se situent à l'intérieur d'une fourchette valide de valeurs. Par exemple, une vérification de l'étendue pourrait être effectuée pour l'âge déclaré par le répondant afin de vérifier s'il se situe entre 0 et 121 ans.

5.2.1 Recodage

Pour tous les enregistrements où des valeurs sont manquantes (aucune réponse fournie), un code de non-réponse ou « non déclaré » (9, 99, 999, etc.) a été attribué à la question.

À l'étape du recodage, la valeur 7 a été attribuée aux réponses « Ne sait pas ». Les questions pour lesquelles on demandait de cocher toutes les réponses qui s'appliquent ont été modifiées afin de créer des réponses dichotomiques (« Oui » = 1 ou « Non » = 2), et ce, pour chaque catégorie de réponse.

Toutes les réponses comprenant du texte (c'est-à-dire les questions ouvertes) ont été retirées du fichier de données et insérées dans un dossier distinct afin de subir des manipulations supplémentaires, telles que du recodage.

5.2.2 Vérification de l'enchaînement

L'enchaînement des questions est automatiquement intégré à l'application. Par exemple, dans la section « Identification des problèmes », on demandait aux répondants de préciser les types de problèmes qu'ils avaient connus au cours des trois années précédentes. L'enchaînement des questions variait en fonction de leurs réponses.

Des questions supplémentaires étaient posées aux répondants ayant sélectionné certains types de problèmes; les autres répondants n'ont pas eu à répondre à ces questions.

La vérification de l'enchaînement vise à déterminer si le répondant a répondu à des questions ne s'appliquant pas à lui et auxquelles il n'aurait donc pas dû répondre. Dans ces cas, une vérification par ordinateur a éliminé automatiquement les données superflues en suivant l'enchaînement des questions dicté par les réponses à des questions antérieures et, parfois, subséquentes.

Les instructions indiquant de « passez à » étant déclenchées par les réponses à des questions portant, par exemple, sur un certain type de problème grave, sont considérées comme des conditions d'enchaînement.

Les réponses aux questions qui ont été sautées en raison de l'enchaînement des questions ont été modifiées, passant de « Non déclaré » (« 9 », « 99 », « 999 », etc.) à « Saut valide » (« 6 », « 96 », « 996 », etc.).

5.2.3 Vérification de la cohérence

La vérification de la cohérence permet de déterminer si les liens entre les questions sont respectés. Cette vérification peut être effectuée en fonction des liens logiques, juridiques, comptables et structurels entre les questions ou les parties d'une question. Par exemple, une vérification de la cohérence au chapitre du lien entre la date de naissance et l'état matrimonial

pourrait être : « une personne de moins de 15 ans peut uniquement avoir "jamais marié" comme état matrimonial ».

Après l'ajout des codes de réserve « Non déclaré » et « Saut valide » lors du traitement, des vérifications de cohérence ont été appliquées aux données, telles que les suivantes :

1. Nombre de personnes de 18 ans et plus dans le ménage (PHH_Q02)

   Si le nombre de personnes de 18 ans et plus dans le ménage (PHH_Q02) est plus élevé que le nombre total de personnes dans le ménage (PHH_Q01), fixer la valeur de la variable PHH_Q02 à « Non déclaré ».

   En outre, si le nombre total de personnes de 18 ans et plus est de 0 (PHH_Q02), fixer la valeur de la variable PHH_Q02 à « Non déclaré ».

2. Coût total approximatif déboursé pour régler le problème (CST_Q20)

   Si au moins un type de coûts associé au problème est déclaré (CST_Q10) et que le coût total indiqué est de 0 $ (CST_Q20), fixer la valeur de la variable CST_Q20 à « Non déclaré ».

3. A reçu des remboursements ou règlements pour le problème (CST_Q30)

   Si un remboursement ou un règlement a été déclaré (CST_Q30 = 1) et que le montant total indiqué pour la variable CST_Q40 est de 0 $, fixer la valeur de la variable CST_Q30 à « 2 » (Non).

5.3 Codage des questions ouvertes

Un total de 28 questions ouvertes tirées de l'ECPJ ont été recodées, comme décrit ci-dessous. Certaines de ces variables ont été utilisées pour créer de nouvelles variables ou ont été enlevées du fichier suite à l'analyse de risque à la confidentialité.

1) Genre du répondant (GDR_S10) : Les réponses ouvertes fournies à la catégorie « Ou veuillez préciser » ont été recodées à « Genre masculin », « Genre féminin », « Diverses identités de genre » ou « Non déclaré ». Pour le FMGD, les réponses de la catégorie « Diverses identités de genre » ont été recodées au hasard à « Genre masculin » ou« Genre féminin ».

2) Orientation sexuelle (SOR_S01) : Les réponses ouvertes fournies à la catégorie « Ou veuillez préciser » ont été recodées. Sur le FMGD, l'orientation sexuelle est présentée en variable dichotomique, soit « hétérosexuelle » ou « une autre orientation sexuelle ».

3) Groupe de population (PG_S05) : Les réponses ouvertes fournies à la catégorie « Autre » ont été recodées à des catégories de réponse existantes pour dériver la variable VISMFLP sur le FMGD. Il s'agit d'une variable dichotomique, « minorité visible » ou « pas une minorité visible ».

4) Conflits ou problèmes (PRI_S05) : Les réponses ouvertes fournies à la catégorie « Autre » ont été recodées selon les catégories de réponse existantes, laissées à « Autre » ou changées en « Non déclaré ». Aux fins de cohérence interne, les changements correspondants ont été apportés aux variables reliées, notamment PRI_Q10 et SERPROP.

5) Pour les variables suivantes, les réponses ouvertes ont été recodées selon l'une des catégories existantes ou laissées dans la catégorie « Autre » :

   - Achat important ou service (CON_S10);

- Employeur ou emploi (EMP_S10);
- Dette ou argent qui vous est dû (DEB_S10);
- Éclatement de la famille (FAM_S10);
- Garde des enfants (CHL_S10);
- Harcèlement (DSH_S10);
- Motif du harcèlement (DSH_S20);
- Nature du harcèlement (DSH_S30);
- Discrimination (DSH_S40);
- Motif de discrimination (DSH_S50);
- Mesure pour régler le problème (AST_S10);
- Type d'avocat contacté (AST_S30);
- Raison pour ne pas avoir contacté un avocat (AST_S40);
- Raisons pour ne pas avoir pris de mesure (AST_S50);
- Aide que pour une partie du problème (LGA_S20);
- Pas d'aide légale (LGA_S30);
- Auriez souhaité avoir pour résoudre problème (STA_S40);
- Coûts associés au problème (CST_S10);
- Difficultés financières (CST_S50);
- Endroit habité quand perdu votre logement (SOC_S30);
- Problèmes connus en raison du problème le plus grave (HLT_S30).

5.4 Création de variables dérivées

Pour faciliter l'analyse des données, un certain nombre de variables ont été dérivées à partir des réponses fournies dans le cadre de l'enquête. Dans ces cas, on a utilisé deux variables ou plus pour en créer une nouvelle.

5.5 Imputation

L'imputation est un processus utilisé pour déterminer et attribuer des valeurs de remplacement aux valeurs manquantes, afin de résoudre les problèmes que suscitent les données manquantes, invalides ou incohérentes. Il faut à cette fin changer certaines des réponses et toutes les valeurs manquantes de l'enregistrement vérifié pour créer un enregistrement plausible et cohérent en soi.

Pour un nombre limité de cas, aucune réponse n'a été fournie à la question portant sur la province de résidence du répondant. Dans ces cas, les valeurs manquantes, soit la province de résidence, ont été imputées dans le fichier de l'échantillon au moyen de SAS. Aucune autre variable du fichier maître n'a été imputée.

5.6 Contrôle de la divulgation

La loi interdit à Statistique Canada de divulguer toute information recueillie qui pourrait dévoiler l'identité d'une personne, d'une entreprise ou d'un organisme sans leur permission ou sans y être autorisé par la Loi sur la statistique. Diverses règles de confidentialité s'appliquent à toutes les données diffusées ou publiées afin d'empêcher la publication ou la divulgation de toute information jugée confidentielle. Au besoin, des données sont supprimées pour empêcher la divulgation directe ou par recoupement de données reconnaissables.

Le fichier de microdonnées ne contiendra aucun identificateur personnel. Les réponses individuelles et les résultats de très petits groupes ne seront jamais diffusés ou communiqués à des partenaires.

En ce qui concerne les données agrégées ou tabulaires, la confidentialité sera maintenue au moyen d'une agrégation ou d'une suppression de données, au besoin.

6.0 Qualité des données

Le présent chapitre permet à l'utilisateur de prendre connaissance des divers facteurs qui influent sur la qualité des données recueillies dans le cadre de l'enquête. On distingue deux principaux types d'erreurs, à savoir les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage.

6.1 Erreurs non dues à l'échantillonnage

Les erreurs non dues à l'échantillonnage sont définies comme des erreurs qui surviennent lors de presque toutes les activités de l'enquête, à l'exception de l'échantillonnage. Elles sont présentes dans les enquêtes avec échantillon et les recensements (contrairement à l'erreur due à l'échantillonnage, qui est présente uniquement dans les enquêtes avec échantillon). Les erreurs non dues à l'échantillonnage peuvent être causées par les sources suivantes : non-réponse, couverture, mesure et traitement des données.

6.1.1 Non-réponse

La non-réponse est une source d'erreurs non dues à l'échantillonnage et d'erreurs dues à l'échantillonnage. La non-réponse découle de l'impossibilité d'obtenir des réponses complètes pour toutes les unités sélectionnées dans l'échantillon. La non-réponse peut entraîner un biais dans les estimations de l'enquête si les non-répondants présentent des caractéristiques très différentes de celles des répondants. La pondération vise à diminuer. Plus le taux de réponse est faible, plus le risque de biais est élevé. Les erreurs non dues à l'échantillonnage sont également une source d'erreur d'échantillonnage. Plus d'information à ce sujet est fournie à la section 6.2 du présent document. Le taux de réponse global était 50,7 %.

6.1.2 Erreurs de couverture

Les erreurs de couverture comprennent les omissions, les ajouts erronés, les répétitions et les erreurs de classification d'unités dans la base de sondage. Les erreurs de couverture peuvent susciter des estimations d'enquête biaisées, et les répercussions peuvent varier parmi les différents sous-groupes de la population.

Les données de l'ECPJ ont été recueillies auprès des personnes âgées de 18 ans et plus vivant dans des logements privés dans les 10 provinces canadiennes, à l'exclusion des personnes vivant dans un logement collectif, une institution ou sur une réserve indienne. Ces exclusions représentent environ 2 % de la population canadienne âgée de 18 ans et plus vivant dans les 10 provinces. Il est souvent impossible de quantifier avec précision les erreurs de couverture. Les sources potentielles d'erreurs de surdénombrement comprennent les répondants qui ont déménagé dans une institution, un logement collectif ou une réserve indienne, mais dont les coordonnées ne reflètent pas ce déménagement.

Un nombre significatif de personnes du suréchantillon autochtone de l'ECPJ n'ont pas indiqué qu'elles s'identifiaient comme Autochtone au moment de l'enquête. À l'inverse, moins de 1 % des personnes de l'échantillon principal ont indiqué qu'elles s'identifiaient comme Autochtone au

moment de l'enquête. Le biais potentiel introduit par ce changement d'identité autochtone a été pris en compte dans le processus de pondération.

6.1.3 Erreurs de mesure

Les erreurs de mesure (parfois appelées erreurs de réponse) désignent la différence entre la réponse inscrite à une question et la valeur réelle. Le répondant, le questionnaire ou la méthode de collecte des données peuvent être à l'origine de ce genre d'erreur. De telles erreurs peuvent être aléatoires ou produire un biais systématique si elles ne sont pas aléatoires.

Il est très couteux de mesurer avec précision l'ampleur de l'erreur de réponse et très peu d'enquêtes procèdent à une évaluation post-enquête. Cependant, les commentaires des intervieweurs et les rapports d'observation fournissent généralement des indices sur les questions que peuvent poser problème (question mal formulée, formation inadéquate de l'intervieweur, mauvaise traduction, jargon technique, absence de texte d'aide, etc.).

Plusieurs processus ont été mis en place pour prévenir et réduire les erreurs de réponse. Ces processus comprennent une revue rigoureuse du questionnaire et sa mise à l'essai à l'aide d'interviews cognitives, le recours à des intervieweurs qualifiés, une formation complète des intervieweurs portant sur les procédures ainsi que le contenu de l'enquête et un suivi continu du processus de collecte des données. En outre, des mesures ont été mises en place lors des étapes du traitement et de la pondération; elles permettent de corriger certaines erreurs, lorsque possible, ou d'en réduire l'effet.

6.1.4 Erreurs de traitement des données

Les erreurs de traitements sont associées aux activités menées une fois que les réponses ont été reçues. Elles comprennent toutes les activités de traitement suivant la collecte et précédant l'estimation. Elles peuvent être aléatoires comme les autres erreurs et accroître ainsi la variance des estimations de l'enquête, ou elles peuvent être systématiques et introduire un biais. Il est difficile d'obtenir des mesures directes des erreurs de traitement, ainsi que de leur incidence sur la qualité des données, puisqu'elles sont souvent confondues avec d'autres types d'erreurs (non-réponse, mesure et couverture).

6.2 Erreurs d'échantillonnage

Une erreur d'échantillonnage est définie comme une erreur découlant de l'estimation d'une caractéristique de la population à partir de la mesure d'une partie de la population plutôt que de l'ensemble de celle-ci. Pour les enquêtes par échantillonnage probabiliste, il existe des méthodes pour calculer les erreurs d'échantillonnage. Ces méthodes découlent directement du plan d'échantillonnage et de la méthode d'estimation utilisée par l'enquête.

La mesure appliquée le plus souvent pour quantifier l'erreur d'échantillonnage est la variance d'échantillonnage. La variance d'échantillonnage détermine à quel point les estimations d'une caractéristique établies à l'aide de différents échantillons de même taille et de même conception diffèrent les unes des autres. Dans le cas des plans d'échantillonnage qui utilisent l'échantillonnage probabiliste, l'ampleur de la variance d'échantillonnage d'une estimation peut être déterminée. Lorsque la variance est relativement grande par rapport à l'estimation, le degré de précision de l'estimation est donc faible.

Les éléments qui ont des répercussions sur l'ampleur de la variance d'échantillonnage comprennent :

1. La variabilité de la caractéristique d'intérêt dans la population : plus la caractéristique dans la population est variable, plus la variance d'échantillonnage est grande.
2. La taille de la population : en général, la taille de la population a des répercussions sur la variance d'échantillonnage seulement pour les populations de petite taille ou de taille moyenne.
3. Le taux de réponse : la variance d'échantillonnage augmente à mesure que la taille de l'échantillon diminue. Étant donné que les non-répondants diminuent en fait la taille de l'échantillon, les non-réponses augmentent la variance d'échantillonnage.
4. Le plan d'échantillonnage et la méthode d'estimation : certains plans d'échantillonnage sont plus efficaces que d'autres parce que, pour la même taille d'échantillon et la même méthode d'estimation, un plan peut donner une variance d'échantillonnage moindre que l'autre.

L'erreur-type d'un estimateur est la racine carrée de sa variance d'échantillonnage. Cette mesure est plus facile à interpréter parce qu'elle donne une indication de l'erreur d'échantillonnage à l'aide de la même échelle que l'estimation, tandis que la variance est basée sur les différences quadratiques.

Le coefficient de variation (c.v.) d'une estimation est une mesure relative de l'erreur d'échantillonnage. Il est défini comme l'estimation de l'erreur type divisée par l'estimation elle-même, habituellement exprimée en pourcentage (10 % au lieu de 0,1). Il est très utile pour mesurer et comparer l'erreur d'échantillonnage de variables quantitatives avec de grandes valeurs positives. Cependant, il n'est pas recommandé pour des estimations telles que les proportions, les estimations des changements ou des différences et les variables qui peuvent avoir des valeurs négatives.

À Statistique Canada, il est considéré comme une pratique exemplaire de faire état de l'erreur d'échantillonnage d'une estimation par l'intermédiaire de son intervalle de confiance de 95 %. L'intervalle de confiance de 95 % d'une estimation signifie que si l'enquête était répétée à maintes reprises, 95 % du temps (ou 19 fois sur 20), l'intervalle de confiance couvrirait la véritable valeur de la population.

7.0 Pondération

Le principe de l'estimation dans un échantillon probabiliste est que chaque unité sélectionnée dans l'échantillon représente, outre elle-même, d'autres unités qui n'ont pas été sélectionnées dans l'échantillon. Par exemple, pour un échantillon aléatoire simple de 100 unités sélectionné dans une population de 5 000 unités, chaque unité dans l'échantillon représente 50 unités dans la population. Le nombre d'unités représentées par une unité de l'échantillon est appelé le poids de sondage de l'unité échantillonnée.

La présente section fournit les détails de la méthode utilisée pour calculer les poids d'échantillonnage pour l'ECPJ.

La pondération pour l'ECPJ a résulté de plusieurs étapes :

1) Calcul des poids de sondage
2) Retrait des unités hors du champ de l'enquête
3) Ajustement pour la non-réponse
4) Traitement des poids influents
5) Calage

Chacune de ces étapes est décrite plus en détails dans les quatre sous-sections suivantes.

7.1 Poids de sondage

Le poids ($W_{1.i}$) initial calculé s'appelle le poids de sondage. Il est égal à l'inverse de la probabilité de sélection de la personne dans chaque province (i représete la province)

$$W_{1.i} = \left(\frac{Nombre\ de\ personnes\ eligible\ dans\ la\ base\ de\ sondage}{Nombre\ d'unités\ échantillonées}\right)$$

Ce poids tient compte de la probabilité qu'une personne ait été sélectionnée pour le recensement complet (détaillé) de 2016. Des poids ont été attribués pour 42 400 unités échantillonnées.

7.2 Suppression des unités hors du champ

Certaines unités ont été identifiées comme hors du champ de l'enquête lors de la collecte (par exemple, personnes ne vivant pas dans une province canadienne ou celles âgées de moins de 18 ans). Dans cette partie, 317 unités ont été identifiées comme hors du champ de l'enquête.

L'étape suivante a été appliquée :

Si unité hors du champ

$$W_{2.i} = 0$$

Sinon

$$W_{2.i} = W_{1.i}$$

7.3 Ajustement pour la non-réponse

Les 42 083 unités restantes ont été séparées en deux groupes, soit 21 170 répondants et 20 913 non-répondants. Étant donné que très peu d'unités ont été identifiées comme étant hors du champ de l'enquête à l'étape précédente, tous les non-répondants ont été considérés comme étant dans le champ. Des données auxiliaires disponibles dans la base de sondage et des paradonnées de collecte ont été utilisées en combinaison pour modéliser la propension à répondre à l'enquête avec une régression logistique. Seules les variables qui étaient prédictives à la fois de la propension à répondre et des variables clés d'intérêt ont été utilisées dans le modèle de régression logistique. Les groupes de réponse homogène (GHR) ont ensuite été construits en utilisant la propension à répondre prédite par le modèle. Les facteurs d'ajustement de non-réponse ont été calculés dans chaque GRH. Les 20 913 non-répondants ont ensuite été retirés des autres ajustements. Les poids ajustés pour la non-réponse ($W_{3,i}$) ont été calculés pour les 21 170 répondants en utilisant la formule suivante :

$$W_{3.i} = W_{2.i} * \left(\frac{\sum W_2 \; de \; répondants + \sum W_2 \; de \; non - répondants}{\sum W_2 \; de \; répondants} \right)$$

Pour les 20 913 non-répondants, $W_{3.i} = 0$.

7.4 Traitement des poids influents

Les unités pour lesquelles l'identité autochtone sur l'ECPJ était différente de celle du recensement complet (détaillé) de 2016 et les unités dont la province sur l'ECPJ était différente de celle du fichier d'échantillon sont considérées comme des unités migrantes (stratum jumpers). Pour certaines de ces unités, les poids dans la strate finale étaient significativement plus grands que pour les unités non-migrantes (non-stratum jumpers) de la même strate. Par exemple, un individu dont la province de résidence sur le recensement de 2016 était l'Ontario mais qui avait depuis déménagé à l'Île-du-Prince-Édouard, aurait un poids beaucoup plus grand que les individus qui ont indiqué résider à l'Île-du-Prince-Édouard lors du recensement de 2016 et de l'ECPJ. Ces unités sont considérées comme migrantes (stratum jumpers). Pour les unités dont les poids étaient plus grands d'au moins trois écarts-types à la moyenne de leur strate, leurs poids ont été réduits pour être égaux au poids maximal des unités non-migrantes (non-stratum jumpers) dans leur strate finale. Au total, les poids ont été ajustés de cette manière pour 41 unités.

7.5 Ajustement aux totaux externes connus

Les poids de sondage ont ensuite été ajustés afin de rendre les estimations de population cohérentes avec les totaux externes connus pour les personnes de 18 ans et plus. Ce processus s'appelle la post-stratification. Les totaux de contrôle externes suivants ont été utilisés :

1) Totaux de population pour chaque province*sexe*group d'âge.
2) Totaux de population par identité autochtone.

Les totaux de population pour la province*sexe*âge ont été obtenus à partir des estimations démographiques, et les totaux de population par identité autochtone ont été estimés à l'aide du recensement détaillé de 2016. Les poids de personne obtenus après cette étape représentent le poids final au niveau de la personne qui est disponible sur le fichier de microdonnées. La somme des poids finaux pour les 21 170 enregistrements incluse dans le fichier final représente l'estimation de la population cible de l'ECPJ.

8.0 Lignes directrices pour la totalisation, l'analyse et la diffusion de données

Le présent chapitre donne un aperçu des lignes directrices recommandées que doivent respecter les utilisateurs qui totalisent, analysent, publient ou diffusent des données calculées à partir des fichiers de microdonnées de l'enquête. Ces lignes directrices devraient permettre aux utilisateurs de microdonnées de fournir les mêmes chiffres que ceux produits par Statistique Canada, tout en étant en mesure d'élaborer des données actuellement inédites qui sont conformes aux lignes directrices établies par l'organisme.

8.1 Lignes directrices pour l'arrondissement

Afin que les estimations, qui sont calculées à partir des fichiers de microdonnées, correspondent à celles produites par Statistique Canada, on conseille vivement aux utilisateurs de respecter les lignes directrices suivantes en ce qui concerne l'arrondissement de telles estimations :

a) Les estimations dans le corps principal d'un tableau statistique doivent être arrondies à l'aide de la technique d'arrondissement ordinaire. Selon cette technique, si le premier ou le seul chiffre à supprimer se situe entre 0 et 4, le dernier chiffre à conserver ne change pas. Si le premier ou le seul chiffre à supprimer se situe entre 5 et 9, le dernier chiffre à conserver est augmenté de 1.

b) Les totaux partiels marginaux et les totaux marginaux des tableaux statistiques doivent être calculés à partir de leurs composantes non arrondies correspondantes, puis arrondis à l'aide de la technique d'arrondissement ordinaire. Les moyennes, les taux, les pourcentages, les proportions et les ratios doivent être calculés à partir de composantes non arrondies (c'est-à-dire des numérateurs ou des dénominateurs), puis arrondis à l'aide de la technique d'arrondissement ordinaire. Les sommes et les différences doivent être calculées à partir des composantes correspondantes non arrondies, puis arrondies à l'aide de la technique d'arrondissement ordinaire.

c) Dans les cas où, en raison de limitations d'ordre technique ou autres, une technique d'arrondissement autre que la technique ordinaire est utilisée produisant des estimations à être publiées ou autrement diffusées qui diffèrent des estimations correspondantes publiées par Statistique Canada, nous conseillons vivement aux utilisateurs d'indiquer la raison de ces différences dans le ou les documents à publier ou à diffuser.

d) En aucun cas les utilisateurs ne doivent publier ou diffuser des estimations non arrondies. Les estimations non arrondies laissent entendre qu'elles sont plus précises qu'elles ne le sont en réalité.

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

L'ECPJ utilise un plan d'échantillonnage et une méthode d'estimation complexes, et par conséquent, les poids d'enquête ne sont pas égaux pour toutes les unités échantillonnées. Lorsqu'ils produisent des estimations simples, y compris des tableaux statistiques ordinaires, les utilisateurs **doivent** appliquer les poids d'enquête appropriés. Sinon, les estimations calculées à partir des fichiers de microdonnées ne peuvent être considérées comme représentatives de la population observée et ne correspondront pas à celles de Statistique Canada.

8.3 Lignes directrices relatives à la qualité pour la diffusion

Avant de diffuser et/ou de publier toute estimation, les analystes doivent prendre en considération le niveau de qualité de l'estimation. Tandis que les erreurs d'échantillonnage et non dues à l'échantillonnage influencent la qualité des données, la présente section porte sur la qualité en ce qui a trait aux erreurs d'échantillonnage. À Statistique Canada, il est considéré comme une pratique exemplaire de faire état de l'erreur d'échantillonnage d'une estimation par l'intermédiaire de son intervalle de confiance de 95 %. L'intervalle de confiance doit être publié avec l'estimation, dans le même tableau que celle-ci. En plus des intervalles de confiance, les estimations sont classées dans l'une des trois catégories de diffusion suivantes :

Catégorie A
Les estimations et intervalles de confiance peuvent être publiés sans mise en garde. Les utilisateurs des données devraient utiliser l'intervalle de confiance de 95 % pour déterminer si la qualité de l'estimation est suffisante. Notez que le 'A' n'est pas un indicateur de qualité, il ne devrait pas être diffusé.

Catégorie E
Les estimations et les intervalles de confiance doivent être signalés par la lettre E (ou un quelconque identificateur similaire) et accompagnés d'un avertissement d'utiliser les estimations avec prudence. Les utilisateurs de données devraient utiliser l'intervalle de confiance de 95% afin de déterminer si la qualité de l'estimation est suffisante.

Catégorie F
Les estimations et les intervalles de confiance ne sont pas recommandés pour la diffusion. Ils sont jugés de si piètre qualité, qu'ils ne se portent à aucune utilisation; ils sont très instables, ce qui les rend peu fiables et potentiellement trompeurs. Si les analystes insistent pour publier des estimations de piètre qualité, même après avoir été informés de leur exactitude, ces estimations doivent être accompagnées d'un avis de non-responsabilité. Les analystes doivent tenir compte des mises en garde reçues et s'engager à ne pas diffuser, présenter, ni déclarer les estimations, directement ou indirectement, sans cet avis de non-responsabilité. Ces estimations doivent être signalées par la lettre F (ou un quelconque identificateur semblable) et la mise en garde suivante doit accompagner les estimations et intervalles de confiance :
« Nous informons l'utilisateur que ces estimations et intervalles de confiance (désignés par la lettre F) ne respectent pas les normes de qualité de Statistique Canada. Les conclusions basées sur ces données ne seront pas fiables et peuvent être invalides.»

Les règles afin d'assigner une estimation à une catégorie de diffusion dépendent du type d'estimation.

Règles de diffusion pour les estimations de proportions et de comptes
Les estimations de proportions et de comptes sont dérivées à partir de variables binaires. Les estimations de comptes sont des estimations du nombre total de personnes/ménages avec une caractéristique d'intérêt; en d'autres termes, elles représentent une somme pondérée d'une variable binaire (ex., nombre estimé d'immigrants). Les estimations de proportions sont des estimations de proportions de personnes /ménages avec une caractéristique d'intérêt (ex., proportion estimée d'immigrants dans la population générale). Les estimations de comptes et de proportions peuvent aussi être dérivées à partir de variables catégorielles: c'est-à-dire l'estimation du nombre ou de la proportion des personnes/ménages qui appartiennent à une catégorie.

Les règles de diffusion pour les estimations de proportions et de comptes sont basées sur la taille d'échantillon. Les tableaux 1, 2 et 3 fournissent les règles de diffusion pour l'ECPJ. Les règles du tableau 2 sont utilisées lorsque le domaine d'intérêt est au niveau de la province ou inférieur (sauf les estimations pour les personnes Autochtones); en d'autres termes, tous les répondants qui contribuent à l'estimation appartiennent au même province. Les règles du tableau 3 sont utilisées

lorsque les estimations sont produites au niveau de l'identité Autochtone. Autrement, les règles du tableau 1 sont utilisées.

Tableau 1: Règles générales pour les proportions et les comptes

Taille d'échantillon (n)	**Catégorie de Diffusion**	**Action**
n > 130	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
65 ≤ n ≤ 130	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC
n < 65	F	Supprimer l'estimation et son IC pour des raisons de qualité

Tableau 2: Règles pour les proportions et les comptes pour des estimations au niveau de la province ou inférieur

Taille d'échantillon (n)	**Catégorie de Diffusion**	**Action**
n > 75	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
37 ≤ n ≤ 75	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC
n < 37	F	Supprimer l'estimation et son IC pour des raisons de qualité

Tableau 3: Règles pour les proportions et les comptes pour des estimations au niveau de l'identité Autochtone

Taille d'échantillon (n)	**Catégorie de Diffusion**	**Action**
n > 90	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
45 ≤ n ≤ 90	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC
n < 45	F	Supprimer l'estimation et son IC pour des raisons de qualité

***** Notez que 'A' n'est pas un indicateur de qualité; il ne devrait pas être diffusé avec l'estimation. L'intervalle de confiance de 95% est l'indicateur de qualité.

Pour des estimations de proportions, *n* est défini comme le compte non pondéré du nombre de répondants au dénominateur (et non au numérateur) de la proportion. Pour des estimations de comptes, *n* est défini comme le compte non pondéré de répondants ayant des valeurs non nulles contribuant à l'estimation.

Règles de diffusion pour les moyennes et les totaux de variables quantitatives

Les règles de diffusion pour des estimations de moyennes et totaux de variables quantitatives sont basées sur la taille d'échantillon et sur le CV de l'estimation. Les tableaux 4, 5 and 6 fournissent les règles de diffusion pour l'ECPJ. Les règles du tableau 5.4 sont utilisées lorsque le domaine d'intérêt est au niveau de la province ou inférieur (sauf les estimations pour les personnes Autochtones); en d'autres termes, tous les répondants contribuant à l'estimation appartiennent au même province. Les règles du tableau 5 sont utilisées lorsque les estimations sont produites au niveau de l'identité Autochtone. Autrement, les règles du tableau 6 sont utilisées.

Tableau 4: Règles générales pour les moyennes et totaux

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n>130 et CV≤25%	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
Autrement	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC
n<65 ou CV>50%	F	Supprimer l'estimation et son IC pour des raisons de qualité

Tableau 5: Règles pour les moyennes et totaux pour des estimations au niveau de la province ou inférieur

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n>75 et CV≤25%	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
Autrement	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC
n<37 ou CV>50%	F	Supprimer l'estimation et son IC pour des raisons de qualité

Tableau 6: Règles pour les moyennes et totaux pour des estimations au niveau de l'identité Autochtone

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n>90 et CV≤25%	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
Autrement	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC
n<45 ou CV>50%	F	Supprimer l'estimation et son IC pour des raisons de qualité

* Notez que 'A' n'est pas un indicateur de qualité; il ne devrait pas être diffusé avec l'estimation. L'intervalle de confiance de 95% est l'indicateur de qualité.

Pour des estimations de moyennes, *n* est défini comme le compte non pondéré du nombre de répondants contribuant à l'estimation incluant les valeurs de zéro. Pour des estimations de totaux, *n* est défini comme le compte non pondéré du nombre de répondants ayant des valeurs non nulles contribuant à l'estimation.

Règles de diffusion pour les différences

Afin d'assigner une catégorie de diffusion pour une estimation de différence entre deux estimations, l'analyste doit d'abord déterminer la catégorie de diffusion de chacune des deux estimations en utilisant les règles décrites précédemment. Ensuite, la catégorie de diffusion de l'estimation de la différence ou de l'estimation du changement est assignée à la catégorie de diffusion la plus basse des deux estimations ; ceci peut être spécifié ainsi :

- Si une estimation ou les deux sont de catégorie F, l'estimation de la différence est assignée à la catégorie F et doit être supprimée
- Sinon, si une estimation ou les deux sont de catégorie E, l'estimation de la différence est assignée à la catégorie E
- Si les deux estimations sont de catégorie A, alors l'estimation de la différence est assignée à la catégorie A

Règles additionnelles au sujet des intervalles de confiance

Les règles de diffusion précédentes devraient supprimer la plupart des estimations et intervalles de confiance de piètre qualité. Il y a également deux conditions qui indiquent si un intervalle de confiance est de piètre qualité. Une estimation et son intervalle de confiance devraient être

assignés à la catégorie de diffusion F si l'une ou l'autre des deux conditions suivantes est satisfaite:

- La borne inférieure de l'intervalle de confiance de 95% est égale à la borne supérieure de l'intervalle; en d'autres termes, l'intervalle de confiance est de longueur nulle. (Une exception survient si l'estimation correspond à un total de contrôle pour l'étalonnage.)
- La borne inférieure ou supérieure de l'intervalle de confiance de 95% n'est pas une valeur plausible pour l'estimation. Par exemple, la borne inférieure d'une estimation de proportion est négative.

8.4 Lignes directrices pour l'analyse statistique, l'estimation de la variance et construction d'intervalles de confiance

Afin de mesurer l'erreur d'échantillonnage des estimations, il faut calculer les estimations de la variance et construire les intervalles de confiance. L'ECPJ utilise un plan d'échantillonnage et une méthode d'estimation complexes, de sorte qu'il n'y a pas de formule simple pour calculer les estimations de la variance. Par conséquent, l'enquête utilise une méthode de rééchantillonnage appelée la méthode bootstrap. Un millier d'ensembles de poids bootstrap a été généré. Essentiellement, la variance est estimée en calculant la valeur de l'estimation d'intérêt au moyen de chacun des ensembles de poids bootstrap, puis la variabilité entre ces 1 000 estimations bootstrap est ensuite mesurée.

Progiciels statistiques pour l'analyse statistique et l'estimation de la variance
Il est nécessaire d'utiliser des poids bootstrap pour calculer des estimations de la variance exactes pour cette enquête. Un certain nombre de programmes ou de progiciels statistiques ont été conçus expressément pour analyser des données fondées sur des plans de sondage complexes et permettre d'estimer la variance au moyen de poids de rééchantillonnage, comme les poids bootstrap. SUDAAN, WesVar, STATA ainsi que les versions plus récentes de SAS en sont des exemples.

D'autres progiciels d'analyse statistique standards ou plus vieux, dont SPSS et les versions de SAS antérieures à la version 9.2, n'ont pas de procédure intégrée pour calculer des estimations de la variance à partir de poids bootstrap lorsque des données fondées sur un plan d'échantillonnage complexe sont utilisées. Ces logiciels ne doivent pas être utilisés pour calculer les estimations de la variance, construire des intervalles de confiance ou procéder à des tests statistiques (tests d'hypothèses, analyse de la régression, et ainsi de suite).

Les versions 9.2 et plus récentes de SAS peuvent calculer la variance au moyen de poids bootstrap ainsi que d'autres types de poids de rééchantillonnage comme des poids jackknife et des poids de réplique répétée équilibrée (RRE). Il existe également un certain nombre de procédures, telles que la régression et la régression logistique, qui acceptent les poids de rééchantillonnage. Les intervalles de confiance pour les médianes utilisant des poids de rééchantillonnage ne sont possibles dans SAS qu'à partir de la version 9.3.

Il est à noter que les progiciels qui ne soutiennent pas explicitement les poids bootstrap, mais qui soutiennent la méthode RRE peuvent être utilisés avec des poids bootstrap. Même si les méthodes bootstrap et RRE diffèrent quant à la manière dont les poids de rééchantillonnage sont construits, une fois ces poids produits, les deux méthodes utilisent une formule similaire pour calculer les estimations de la variance. Pour obtenir de plus amples renseignements sur la relation entre la méthode bootstrap et la méthode RRE, veuillez-vous reporter à Phillips (2004).

Facteur multiplicatif
La méthode utilisée pour créer les poids bootstrap pour l'ECPJ comprenait une étape où les poids bootstrap ont été transformés afin d'éliminer les poids négatifs. Les poids bootstrap ainsi

transformés exigent que les estimations de la variance pour l'ECPJ soient multipliés par un facteur de 4. **Il est extrêmement important d'appliquer ce facteur multiplicatif. L'omission de ce facteur entraînerait des résultats et des conclusions incorrects.**

Les progiciels statistiques qui permettent l'utilisateur d'utiliser le BRR avec l'ajustement de Fay peuvent produire des estimations de variance correctes sans avoir besoin d'une étape de multiplication supplémentaire. Le facteur multiplicatif peut être spécifié en utilisant le paramètre Fay : pour certains progiciels (par exemple, SUDAAN), utilisez un paramètre Fay de C, où C est le facteur multiplicatif de la variance. Pour d'autres progiciels (SAS, en particulier), utilisez un facteur de Fay k, où $k = 1 - \sqrt{\frac{1}{C}}$. Pour CLPS, $C = 4$ et $k = 0.5$.

Intervalles de confiance

La méthode la plus couramment utilisée pour construire des intervalles de confiance de 95 % est l'intervalle de Wald, qui est de la forme $\hat{y} \pm 1.96\sqrt{\text{vâr}(\hat{y})}$ pour une estimation $\hat{y}$ avec l'estimation de variance vâr$(\hat{y})$. Les intervalles de Wald sont fondés sur l'hypothèse selon laquelle la répartition d'échantillonnage de $\hat{y}$ est à peu près normale. En ce qui concerne les proportions, il est établi que l'hypothèse de normalité ne tient plus dans le cas des échantillons de petite taille et des proportions situées près de 0 ou de 1. Trois autres méthodes pour construire des intervalles de confiance sont par conséquent recommandées pour les proportions : l'intervalle de Wilson modifié, l'intervalle de Clopper-Pearson modifié et l'intervalle logit (voir Korn et Graubard, 1998 et Liu et Kott, 2009). Il existe des options dans SAS et SUDAAN pour produire des intervalles de confiance à l'aide de ces autres méthodes.

Les exemples ci-dessous montrent comment ces méthodes recommandées pour construire des intervalles de confiance sont précisées pour des proportions dans SAS et SUDAAN, et ils indiquent le facteur Fay nécessaire tel qu'expliqué à la section précédente.

1. SAS, intervalles de confiance de Wilson modifiés :

```
PROC SURVEYFREQ
DATA=…. VARMETHOD=BRR (Fay = 0.5);
WEIGHT WTPP;
REPWEIGHTS WRPP1-WRPP1000;
TABLES …. / CL (TYPE=WILSON ADJUST=NO TRUNCATE=YES)
```

2. SUDAAN, intervalles de confiance de Clopper-Pearson modifiés :

```
PROC CROSSTAB
DATA=…. DESIGN=BRR SMCONF=50;
WEIGHT WTPP;
REPWGT WRPP1-WRPP1000 / ADJFAY=4;
TABLES ...;
```

Rééchelonnage des poids

Comme nous l'avons déjà mentionné, il est recommandé que les utilisateurs utilisent les procédures d'analyse conçues pour analyser des données à partir de plans d'échantillonnage complexes, qui peuvent utiliser les poids pour produire des estimations et les poids bootstrap pour produire des estimations de la variance. Certaines procédures d'analyse qui ne sont pas conçues pour le contexte de l'échantillonnage peuvent permettre l'utilisation des poids (sans poids bootstrap). Cependant, ces procédures peuvent différer dans leur définition des poids, et produire des estimations exactes mais des estimations de la variance dénuées de sens. Pour des analyses telles que la régression linéaire, la régression logistique et l'analyse de variance, le rééchelonnement des poids peut rendre la variance calculée par les progiciels standards plus raisonnable. Les poids pour le domaine d'intérêt devraient être rééchelonnés de sorte que le poids moyen est d'un (1); ceci peut être accompli en divisant chaque poids par le poids moyen global du domaine avant de procéder à l'analyse. Le rééchelonnage rend les estimations de la

variance plus raisonnables, mais celles-ci tiennent compte uniquement de l'inégalité des probabilités de sélection – elles ne tiennent pas compte ni de la stratification ni des grappes du plan d'échantillonnage. Cette approche devrait donc être utilisée qu'en dernier recours, seulement lorsque qu'aucune procédure permettant l'utilisation des poids bootstrap n'est disponible; les utilisateurs sont avisés que les résultats sont approximatifs.

Références

Korn, E.L., and Graubard, B.I. (1998). "Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data". *Survey Methodology*, 24, 193-201.

Liu, Y.K. and Kott, P.S. (2009). "Evaluating Alternative One-Sided Coverage Intervals for a Proportion". *Journal of Official Statistics*, Vol. 25, No. 4, 569-588.

Phillips, O. (2004). "Using bootstrap weights with WesVar and SUDAAN" (Catalogue no. 12-002-X20040027032) in The Research Data Centres Information and Technical Bulletin, Chronological index, Fall 2004, vol.1 no. 2 Statistics Canada, Catalogue no. 12-002-XIE.