

# **Guide de l'utilisateur des microdonnées**

## **Enquête canadienne sur l'alcool et les drogues de 2019**

**Juin à décembre 2019**



Statistics  
Canada

Statistique  
Canada

**Canada**

## Table des matières

|     |   |    |
|-----|---|----|
| 1.0 | INTRODUCTION .....  | 2  |
| 2.0 | CONTEXTE .....  | 3  |
| 3.0 | OBJECTIFS.....  | 4  |
| 4.0 | MÉTHODOLOGIE DE L'ENQUÊTE .....   | 5  |
| 4.1 | Population visée et base de sondage .....   | 5  |
| 4.2 | Plan de sondage et répartition de l'échantillon.....                                | 5  |
| 4.3 | Pondération .....   | 6  |
| 5.0 | COLLECTE DES DONNÉES.....   | 7  |
| 5.1 | Conception du questionnaire.....  | 7  |
| 5.2 | Collecte et vérification des données .....  | 7  |
| 6.0 | TRAITEMENT DES DONNÉES.....   | 8  |
| 6.1 | Saisie des données .....  | 8  |
| 6.2 | Vérification .....  | 8  |
| 6.3 | Création de variables dérivées .....  | 8  |
| 6.4 | Suppression de renseignements confidentiels pour le FMGD .....                      | 8  |
| 7.0 | QUALITÉ DES DONNÉES.....  | 10 |
| 7.1 | Taux de réponse .....   | 10 |
| 7.2 | Erreurs d'enquête .....   | 11 |
| 7.3 | Non-réponse totale.....   | 11 |
| 7.4 | Non-réponse partielle .....   | 11 |
| 7.5 | Couverture.....   | 11 |
| 7.6 | Mesure de l'erreur d'échantillonnage .....  | 12 |
| 8.0 | LIGNES DIRECTRICES POUR LA TOTALISATION, L'ANALYSE ET LA DIFFUSION DE DONNEES ..... | 13 |
| 8.1 | Lignes directrices pour l'arrondissement d'estimations .....                        | 13 |

|   |   |           |
|---|---|-----------|
| 8.2   | Lignes directrices pour la pondération de l'échantillon en vue de la totalisation ..... | 14        |
| 8.3   | Définitions de types d'estimations : catégorielles et quantitatives .....               | 14        |
| 8.3.1   | Estimations catégorielles .....   | 14        |
| 8.3.2   | Estimations quantitatives .....   | 14        |
| 8.3.3   | Totalisation d'estimations catégorielles .....  | 15        |
| 8.3.4   | Totalisation d'estimations quantitatives .....  | 15        |
| 8.4   | Lignes directrices pour l'analyse statistique .....                                     | 16        |
| 8.5   | Lignes directrices de diffusion .....   | 17        |
| 8.5.1   | Lignes directrices de diffusion basées sur la qualité .....                             | 17        |
| 8.5.2   | Lignes directrices de diffusion basées sur la confidentialité .....                     | 19        |
| 9.0   | PONDÉRATION .....   | 20        |
| 9.1   | Procédures de pondération .....   | 20        |
| 10.0  | AUTRE DOCUMENTATION .....   | 22        |
| <b>ANNEXE A – ESTIMATION DE LA VARIANCE ET CONSTRUCTION DES<br/>INTERVALLES DE CONFIANCE.....</b> |   | <b>23</b> |

## **1.0 Introduction**

L'Enquête canadienne sur l'alcool et les drogues a été menée par Statistique Canada entre juin et décembre 2019 avec l'appui et la collaboration de Santé Canada. Le présent document a été produit pour faciliter la manipulation des fichiers de microdonnées (fichier maître et FMGD - Fichier de microdonnées à grande diffusion) des résultats de l'enquête.

Toutes questions au sujet des ensembles de données ou de leur utilisation doivent être adressées à :

### Statistique Canada

Services 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

Télécopieur : 613-951-4527

Courriel: [statcan.csdidclientservice-ciddsservicealaclientele.statcan@canada.ca](mailto:statcan.csdidclientservice-ciddsservicealaclientele.statcan@canada.ca)

### Santé Canada

Direction des substances contrôlées

Direction générale des substances contrôlées et du cannabis

Ottawa (Ontario) K1A 0K9

Courriel : [hc.odss-bssd.sc@canada.ca](mailto:hc.odss-bssd.sc@canada.ca)

## **2.0 Contexte**

De 1999 à 2012, les données sur le tabagisme ont été collectées chaque année dans le cadre de l'enquête de surveillance de l'usage du tabac au Canada (ESUTC). En 2013, 2015 et 2017, Statistique Canada menait l'Enquête canadienne sur le tabac, l'alcool et les drogues (ECTAD). Cette enquête permettait de recueillir des données sur le tabac, mais aussi sur l'alcool et les drogues. Pour la première fois en 2019, l'Enquête canadienne sur l'alcool et les drogues (ECAD) a été menée et a permis de collecter des données principalement sur l'alcool et les drogues. Indépendamment, en 2019, l'Enquête canadienne sur le tabac et la nicotine (ECTN) a été menée afin de collecter des données sur le tabac et la nicotine.

### **3.0 Objectifs**

L'objectif principal de cette enquête est de recueillir des renseignements sur les Canadiens au sujet de leur consommation d'alcool et de drogues. Santé Canada et d'autres organismes utiliseront ces renseignements afin de suivre l'évolution de la consommation d'alcool et de drogues.

Les objectifs complémentaires de l'Enquête canadienne sur l'alcool et les drogues (ECAD) sont les suivants : mesurer la fréquence de la consommation d'alcool, mesurer la fréquence de la consommation de cannabis, mesurer la fréquence de la consommation d'autres drogues et mesurer les méfaits possibles de la consommation d'alcool, de cannabis et des autres drogues.

L'ECAD est la seule enquête de Statistique Canada qui répond au besoin de Santé Canada d'avoir de l'information continue et détaillée sur la prévalence de l'usage de drogues et la consommation d'alcool selon la province, le sexe ou le groupe d'âge, pour les groupes d'âge de 15 à 19 ans, de 20 à 24 ans et de 25 ans et plus.

## **4.0 Méthodologie de l'enquête**

L'Enquête canadienne sur l'alcool et les drogues (ECAD) a été menée entre le 10 juin et le 31 décembre 2019 par questionnaire électronique et, en cas de non réponse, par un suivi téléphonique.

### **4.1 Population visée et base de sondage**

La population cible de l'ECAD était composée de toutes les personnes âgées de 15 ans et plus vivant au Canada, à l'exception des personnes suivantes :

- 1) les résidents du Yukon, des Territoires du Nord-Ouest et du Nunavut;
- 2) les résidents à temps plein d'un établissement institutionnel;
- 3) les résidents des réserves amérindiennes.

L'enquête a utilisé le fichier de l'univers des logements (FUL), un fichier produit par Statistique Canada, comme base de sondage. Ceci a été fait dans le but de produire des estimations de qualité au niveau provincial et des différents groupes d'âge (au niveau du Canada), et de faciliter un premier contact par envoi postal pour l'invitation à compléter le questionnaire de façon électronique. Cette base de sondage permet d'avoir jusqu'à trois numéros de téléphone pour permettre le suivi téléphonique avec un ménage, incluant des numéros de téléphone filaire et cellulaire. Un processus de nettoyage de l'échantillon visant à éliminer les numéros de téléphone qui n'étaient pas en service ou inconnus a été exécuté avant d'envoyer l'échantillon à l'équipe responsable de la collecte.

Puisque l'enquête a été menée à partir d'un échantillon d'adresses, presque tous les ménages ont pu être contactés par la poste. Les logements ayant été identifiés comme étant vacants au moment de créer la base de sondage ont été exclus. Les logements qui n'avaient ni adresse postale, ni numéro de téléphone associé ont également été exclus de la base de sondage, étant donné qu'ils ne pouvaient pas être contactés par aucun des modes de collecte de l'enquête. Toutefois, les estimations de l'enquête ont été pondérées afin d'inclure les personnes vivant dans de tels logements.

### **4.2 Plan de sondage et répartition de l'échantillon**

Le plan de sondage de l'ECAD 2019 était un échantillon aléatoire stratifié à deux phases. Les provinces formaient les strates. Durant la première phase, les ménages étaient sélectionnés de façon aléatoire, et durant la deuxième phase, une personne était sélectionnée à l'intérieur du ménage selon la méthode de sélection basée sur l'âge. L'algorithme de sélection était basé sur le nombre de membres éligibles dans le ménage et sur l'âge ordonné de chacun de ses membres.

Une lettre a été envoyée au ménage sélectionné et un seul membre du ménage était choisi, via les instructions fournies dans la lettre, pour compléter le questionnaire électronique. La personne sélectionnée était invitée à compléter le questionnaire en y accédant en ligne et en entrant un code d'accès sécurisé (CAS) fourni dans la lettre. Les interviews par procuration n'étaient pas acceptées.

La sélection basée sur l'âge a également été utilisée pour les répondants ITAO (interviews téléphoniques assistées par ordinateur). La sélection était faite avec l'intervieweur. Les instructions de la lettre, de même que la sélection effectuée avec l'interviewer, étaient cohérentes pour un même ménage échantillonné, afin de s'assurer que la même personne était sélectionnée pour participer à l'enquête pour un ménage donné, peu importe le mode de collecte utilisé pour répondre au questionnaire.

Une allocation de Kish a été utilisée pour allouer l'échantillon de sorte que les cibles de qualité soient respectées pour plusieurs domaines d'intérêt. La taille initiale a été déterminée en supposant un taux de réponse global de 50% et un effet de plan de 1.5. Il a été déterminé qu'un échantillon de 22,000 ménages était nécessaire afin de produire des estimations de qualité au niveau provincial. L'allocation de l'échantillon peut être trouvée à la section 7.0 *Qualité des données*.

### **4.3 Pondération**

Le principe qui sous-tend une estimation pour un échantillon probabiliste veut que chaque personne incluse dans l'échantillon « représente », en plus d'elle-même, plusieurs autres personnes qui en sont exclues. Par exemple, dans un échantillon aléatoire simple de 2 % de la population, chaque personne incluse dans l'échantillon représente 50 membres de la population.

La phase de la pondération est une étape où l'on calcule ce nombre (ou poids) pour chaque enregistrement. Ce poids, qui figure dans le fichier de microdonnées, doit être utilisé afin de calculer des estimations significatives à partir des données de l'enquête. Si, par exemple, le nombre de personnes au Canada qui consomment de l'alcool tous les jours doit être estimé, cette opération s'effectue en sélectionnant les enregistrements renvoyant aux personnes incluses à l'intérieur de l'échantillon qui présentent cette caractéristique (ALC\_Q15 = 1) et en additionnant les poids inscrits dans ces enregistrements. Le chapitre 10.0 — *Pondération* renferme des détails au sujet de la méthode utilisée pour calculer ces poids.



## **5.0 Collecte des données**

### **5.1 Conception du questionnaire**

Le questionnaire de l'ECAD de 2019 se fonde en partie sur celui de l'Enquête canadienne sur le tabac, l'alcool et les drogues (ECTAD) 2017 et des versions précédentes. Le questionnaire comporte un pourcentage élevé de nouvelles questions. Les sections sexe, genre et âge (AGS), expériences maternelles avec l'utilisation de cannabis et d'alcool (MEX), Spice (SPI), Kratom (KRT), Méphédron (MEP), BZP ou TFMPP (BZP), consommation de drogues injectables (IDU), surdose (OD) et traitement (TT) sont nouveaux. Plusieurs sections déjà existantes dans l'ECTAD ont aussi été modifiées en changeant l'ordre des questions, en ajoutant quelques questions, en séparant des questions en deux différentes, etc.

C'était la première fois que CADS 2019 utilisait un questionnaire électronique à l'instar de l'utilisation exclusive d'interviews téléphoniques assistées par ordinateur (ITAO).

Des spécifications définissant les limites valides et garantissant la cohérence d'une question à l'autre ont été intégrées dans l'application du questionnaire électronique dans la mesure du possible. Des contrôles de cohérences supplémentaires ont été réalisés durant la phase de traitement des données.

### **5.2 Collecte et vérification des données**

La collecte de données a été menée chaque mois, de juin à décembre 2019. La collecte s'est divisée en 2 vagues. La première vague s'étendait du 10 juin au 22 septembre. La deuxième vague, quant à elle, était du 23 septembre au 31 décembre. Lors de la première vague, une première lettre d'invitation a été envoyée, suivi de jusqu'à 4 lettres de rappels en cas de non-réponse de la part du ménage. Pour la deuxième vague, une première lettre d'invitation a aussi été envoyée, suivi des 4 lettres de rappels.

Des enchaînements valides et des messages de validation apparaissaient tout au long du questionnaire électronique. Des contrôles intégrés à l'application garantissaient la cohérence des réponses, repéraient et corrigeaient les valeurs aberrantes et déterminaient à qui étaient posées certaines questions. Ainsi, à la fin du processus de collecte, les données étaient déjà passablement « épurées ».

Après les 4 lettres de rappel, tous les cas de non-réponse ont été distribués dans deux bureaux régionaux de Statistique Canada pour un suivi téléphonique (ITAO). La charge de travail et les intervieweurs de chaque bureau étaient supervisés par un gestionnaire de projet. L'ordonnanceur automatique utilisé dans le système d'ITAO garantissait que les cas étaient attribués au hasard aux intervieweurs et que les appels se faisaient à différents moments de la journée pendant des jours différents de la semaine pour maximiser la probabilité de prise de contact.

## **6.0 Traitement des données**

Par le passé, dans le cas de l'Enquête canadienne sur le tabac, l'alcool et les drogues (ECTAD), les principaux produits étaient deux fichiers de microdonnées « épurés », le premier contenant des renseignements sur les ménages; le deuxième contenant des renseignements sur les personnes ainsi qu'un ensemble équivalent de fichiers de microdonnées à grande diffusion (FMGD). À présent, dans l'ECAD, un seul fichier maître et un seul fichier PUMF sont créés. Ce chapitre présente un résumé des phases de traitement inhérentes à la production de ce fichier.

### **6.1 Saisie des données**

Puisque les données ont été recueillies à l'aide d'une application électronique et d'un suivi téléphonique (ITAO - interview téléphonique assistée par ordinateur) un système de collecte des données distinct n'était pas nécessaire.

### **6.2 Vérification**

Le premier type d'erreurs traitées avait trait à l'absence d'information « non déclaré » dans des questions posées au répondant. Par conséquent, un code de non-réponse ou « non déclaré » a été attribué à certaines questions subséquentes.

De plus, quelques incohérences au niveau des réponses fournies dans l'enquête par les répondants ont été corrigées. Parfois, les répondants ont affirmé une information et plus loin dans le questionnaire, mentionné le contraire. En collaboration avec le client de l'enquête, une série de spécifications ont été écrites afin de remédier à ces situations d'incohérence.

### **6.3 Création de variables dérivées**

Un certain nombre de variables incluses dans le fichier de microdonnées ont été calculées en combinant des données élémentaires afin d'en faciliter leur analyse. Le statut de consommateur d'alcool ainsi que le statut de consommateur de cannabis sont des exemples de variables dérivées. La caractéristique rurale ou urbaine de la communauté où habite le répondant (DVURBAN) a été dérivé à partir du code postal.

### **6.4 Suppression de renseignements confidentiels pour le FMGD**

Pour l'ECAD, un fichier de microdonnées à grande diffusion (FMGD) est disponible. Il convient de souligner que le fichier de microdonnées à grande diffusion diffère du fichier « maître » de l'enquête que conservent Statistique Canada. Ces différences sont le résultat de mesures prises pour protéger l'anonymat des répondants à une enquête. Comme le FMGD est accessible publiquement à un large éventail d'utilisateurs, il est essentiel que des mesures supplémentaires soient prises pour garantir que les données des répondants dans le FMGD soient sécuritaires.

Ces mesures supplémentaires comprennent la limitation de la quantité d'information sur la famille et le ménage, l'agrégation des codes et le plafonnement de certaines variables, ou la suppression ou la perturbation des réponses pour certains répondants. Les utilisateurs ayant besoin d'avoir accès à de l'information exclue du FMGD peuvent accéder au fichier maître de microdonnées dans les centres de données de recherche à Statistique Canada ou peuvent acheter des totalisations personnalisées. Les estimations produites par les totalisations personnalisées seront communiquées à l'utilisateur, sous réserve du respect des lignes directrices pour l'analyse et la diffusion dont le *chapitre 9.0 — Lignes directrices pour la totalisation, l'analyse et la diffusion de données* de ce document fournit un aperçu.

Toutes les variables du fichier maître ont été examinées en termes de risque de divulgation résiduelle. Suite à l'analyse et aux mesures correctives prises, le FMGD contient 10 293

enregistrements de répondants et 387 variables. Vous trouverez plus d'informations dans le dictionnaire de données du FMGD.

## 7.0 Qualité des données

### 7.1 Taux de réponse

Pour l'Enquête canadienne sur l'alcool et les drogues (ECAD), un taux de réponse global est calculé comme suit.

Le taux de réponse est la proportion d'enregistrements de personnes sélectionnées, dans le champ de l'enquête, qui contiennent des données valides. Comme une seule personne est sélectionnée par ménage, le taux de réponse ménage est le même que le taux de réponse personne. Par conséquent, seul le terme "taux de réponse" est utilisé. La proportion d'unités répondantes est ajustée par un facteur estimant la proportion de ménages dans le champ de l'enquête. Ce facteur d'ajustement est obtenu en divisant le nombre de ménages de la base de sondage par le nombre de ménages selon les projections démographiques.

$$\frac{\text{nombre de personnes ayant des données valides}}{\text{nombre de ménages sélectionnés}} * \text{facteur d'ajustement}$$

Une **personne répondante** (avec données valides) est définie selon les caractéristiques suivantes :

- la personne a complété la sélection selon l'âge;
- la personne a répondu aux questions sur l'âge, le genre, la taille du ménage et la composition du ménage;
- la personne sélectionnée a répondu à au moins deux questions clés concernant la consommation d'alcool.

**Tableau 1 : Taux de réponse selon la province**

| Province                | Nombre de personnes sélectionnées | Nombre de personnes répondantes | Facteur d'ajustement pour estimer le nombre de ménages dans le champ de l'enquête | Taux de réponse (%) |
|-------------------------|-----------------------------------|---------------------------------|---|---------------------|
| Terre-Neuve-et-Labrador | 1 935                             | 744                             | 1.28  | 49.2                |
| Île-du-Prince-Édouard   | 1 915                             | 827                             | 1.24  | 53.3                |
| Nouvelle-Écosse         | 1 923                             | 891                             | 1.17  | 54.4                |
| Nouveau-Brunswick       | 1 910                             | 903                             | 1.17  | 55.2                |
| Québec                  | 2 786                             | 1 432                           | 1.12  | 57.6                |
| Ontario                 | 3 488                             | 1 638                           | 1.05  | 49.4                |
| Manitoba                | 1 940                             | 954                             | 1.09  | 53.5                |
| Saskatchewan            | 1 938                             | 905                             | 1.11  | 51.9                |
| Alberta                 | 2 096                             | 983                             | 1.07  | 50.4                |
| Colombie-Britannique    | 2 185                             | 1 016                           | 1.09  | 50.8                |
| <b>Total</b>            | <b>22 116</b>                     | <b>10 293</b>                   | <b>1.09</b>   | <b>50.7</b>         |

Note : le facteur d'ajustement est arrondi dans le tableau ci-dessus, donc les taux de réponse présentés ne peuvent être reproduits.

## **7.2 Erreurs d'enquête**

Les estimations calculées à partir des données de cette enquête reposent sur un échantillon de ménages. Des estimations légèrement différentes auraient pu être obtenues si un recensement complet avait été effectué en utilisant le même questionnaire et en faisant appel aux mêmes intervieweurs, superviseurs, méthodes de traitement, etc. que ceux utilisés pour mener l'enquête. L'écart entre les estimations découlant de l'échantillon et celles que donnerait un dénombrement complet mené dans des conditions semblables est appelé erreur d'échantillonnage de l'estimation.

Des erreurs qui ne sont pas liées à l'échantillonnage peuvent se produire à presque toutes les étapes des opérations d'enquête. Les intervieweurs peuvent avoir mal compris les instructions, les répondants peuvent se tromper en répondant aux questions, les réponses peuvent être mal saisies sur le questionnaire électronique et des erreurs peuvent survenir lors du traitement et de la totalisation des données. Ces erreurs sont toutes des exemples d'erreurs non dues à l'échantillonnage.

Sur un grand nombre d'observations, les erreurs aléatoires auront peu d'effet sur les estimations calculées à partir de l'enquête. Toutefois, les erreurs systématiques contribueront à biaiser les estimations de l'enquête. Énormément de temps et d'efforts ont été consacrés à réduire les erreurs non dues à l'échantillonnage dans l'enquête. Des mesures d'assurance de la qualité ont été prises à chacune des étapes du cycle de collecte et de traitement des données afin de contrôler la qualité des données. Ces mesures comprennent la mise à l'essai du questionnaire électronique, la formation poussée des intervieweurs concernant les procédures de l'enquête et de l'application de l'interview téléphonique assistée par ordinateur (ITAO), l'observation des intervieweurs en vue de cerner les problèmes liés à la conception du questionnaire ou à une mauvaise compréhension des instructions, et l'évaluation de l'application pour s'assurer que les contrôles des limites, les vérifications et l'enchaînement des questions étaient tous programmés correctement.

## **7.3 Non-réponse totale**

Dans beaucoup d'enquêtes, la non-réponse totale peut être une source importante d'erreurs non dues à l'échantillonnage, selon la mesure dans laquelle les répondants et les non-répondants diffèrent quant aux caractéristiques présentées. S'il y a eu non-réponse totale, c'est parce que l'intervieweur était incapable de communiquer avec le répondant ou que le répondant a refusé de participer à l'enquête. Les cas de non-réponse totale ont été traités en ajustant le poids des ménages qui ont répondu au questionnaire de l'enquête de façon à le contrebalancer pour ceux qui n'ont pas répondu au questionnaire.

## **7.4 Non-réponse partielle**

Dans la plupart des cas, il y a eu non-réponse partielle au questionnaire de l'enquête lorsque le répondant n'a pas compris ou a mal interprété une question, a refusé d'y répondre ou ne pouvait se rappeler de l'information demandée. Des codes dans le fichier de microdonnées indiquent les cas de non-réponse partielle (c.-à-d. Non-réponse).

## **7.5 Couverture**

Comme mentionné à la *section 4.1 — Population visée et base de sondage*, certains ménages au Canada n'ont ni d'adresse postale valide ni de téléphone de ligne terrestre ou de téléphone cellulaire sur la base de sondage utilisée. Les personnes qui vivent dans ces ménages peuvent avoir des caractéristiques uniques qui ne seront pas reflétées dans les estimations de l'enquête. Les utilisateurs devraient faire preuve de prudence lorsqu'ils analysent des sous-groupes de la population dont les caractéristiques peuvent être corrélées au fait de ne pas avoir d'adresse postale valide ni de téléphone de ligne terrestre ou de téléphone cellulaire.

## **7.6 Mesure de l'erreur d'échantillonnage**

Puisqu'il est inévitable que des estimations établies à partir de données d'une enquête-échantillon soient sujettes à une erreur d'échantillonnage, une saine pratique de la statistique exige que les chercheurs fournissent aux utilisateurs une certaine indication de l'importance de cette erreur d'échantillonnage. Cette section de la documentation renferme un aperçu des mesures de l'erreur d'échantillonnage dont Statistique Canada se sert couramment et dont l'organisme conseille vivement aux utilisateurs qui produisent des estimations à partir de ce fichier de microdonnées à employer également.

La base pour mesurer l'importance potentielle des erreurs d'échantillonnage est l'erreur type des estimations calculées à partir des résultats d'une enquête.

Cependant, en raison de la grande diversité des estimations pouvant être produites à partir d'une enquête, l'erreur type d'une estimation est habituellement exprimée en fonction de l'estimation à laquelle elle se rapporte. Une des mesures résultantes souvent utilisée, est appelée coefficient de variation (c.v.) d'une estimation; elle s'obtient en divisant l'erreur type de l'estimation par l'estimation elle-même et s'exprime en pourcentage de l'estimation. Cette mesure de qualité a été utilisée dans les itérations précédentes des enquêtes sur l'alcool, les drogues et le tabac. Ceci étant dit, puisque de très petites proportions sont parfois mesurées dans l'ECAD, il est préférable d'exprimer la qualité des estimations en présentant leur intervalle de confiance. En effet, une petite proportion aura un CV élevé par construction. En contrepartie, sa proportion complémentaire (1-p) aura un petit CV par construction. Une mesure plus adéquate de la qualité dans ce cas est d'observer l'intervalle de confiance. Ceci laisse aussi le soin à l'utilisateur de déterminer si l'estimation présentée est assez précise pour ses besoins.

Nous recommandons d'utiliser l'intervalle de Wilson modifié, l'intervalle de Clopper-Pearson modifié ou l'intervalle logit pour les proportions binomiales (1/0 ; oui/non ; etc.).

La plupart des logiciels statistiques, tels que SAS ou SUDAAN, permettent de produire ces types d'intervalles. Pour de plus amples informations sur la façon de les calculer, veuillez-vous référer à l'Annexe A.

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

Ce chapitre de la documentation renferme un aperçu des lignes directrices que doivent respecter les utilisateurs qui totalisent, analysent, publient ou autrement diffusent des données calculées à partir du fichier de microdonnées de l'enquête. Ces lignes directrices devraient permettre aux utilisateurs de microdonnées de produire les mêmes chiffres que ceux produits par Statistique Canada, tout en étant en mesure d'obtenir des chiffres actuellement inédits de façon conforme à ces lignes directrices établies.

### **8.1 Lignes directrices pour l'arrondissement d'estimations**

Afin que les estimations qui sont destinées à la publication ou à toute autre forme de diffusion qui sont calculées à partir de ce fichier de microdonnées correspondent à celles produites par Statistique Canada, nous conseillons vivement aux utilisateurs de respecter les lignes directrices qui suivent en ce qui concerne l'arrondissement de telles estimations :

- a) Les estimations dans le corps principal d'un tableau statistique doivent être arrondies à la centaine près à l'aide de la technique d'arrondissement normale. 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. Par exemple, selon la technique d'arrondissement normale à la centaine près, si les deux derniers chiffres se situent entre 00 et 49, ils sont remplacés par 00 et le chiffre précédent (le chiffre des centaines) reste inchangé. Si les deux derniers chiffres se situent entre 50 et 99, ils sont remplacés par 00 et le chiffre précédent est augmenté de 1.
- b) Les sous-totaux marginaux et les totaux des tableaux statistiques doivent être calculés à partir de leurs composantes non arrondies correspondantes, puis être arrondis à leur tour à la centaine près à l'aide de la technique d'arrondissement normale.
- c) Les moyennes, les proportions, les taux et les pourcentages doivent être calculés à partir de composantes non arrondies (c.-à-d. des numérateurs et/ou des dénominateurs), puis être arrondis à leur tour à une décimale à l'aide de la technique d'arrondissement normale. Dans le cas d'un arrondissement normal à un seul chiffre, si le dernier 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.
- d) Les sommes et les différences d'agrégats (ou de rapports) doivent être calculées à partir de leurs composantes non arrondies correspondantes, puis être arrondies à leur tour à la centaine près (ou à la décimale près) à l'aide de la technique d'arrondissement normale.
- e) Dans les cas, où, en raison de limites d'ordre technique ou de toutes autres limites, une technique d'arrondissement autre que la technique normale est utilisée produisant des estimations à être publiées ou autrement diffusées différentes 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.
- f) En aucun cas, les utilisateurs ne doivent publier ou autrement diffuser des estimations non arrondies. Des estimations non arrondies laissent entendre qu'elles sont plus précises qu'elles le sont en réalité.

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

Le plan d'échantillonnage utilisé pour l'ECAD n'était pas autopondéré. Lorsqu'ils produisent des estimations simples, y compris des tableaux statistiques ordinaires, les utilisateurs doivent appliquer le poids d'enquête.

Si les poids ne sont pas utilisés, les estimations calculées à partir du fichier de microdonnées ne peuvent être considérées comme représentatives de la population visée par l'enquête et ne correspondront pas à celles produites par Statistique Canada.

Les utilisateurs devraient également prendre note que certains progiciels pourraient peut-être ne pas permettre la production d'estimations correspondant exactement à celles qu'offrent Statistique Canada, en raison de la façon dont les poids sont appliqués.

## **8.3 Définitions de types d'estimations : catégorielles et quantitatives**

Avant de discuter de la façon dont on peut totaliser et analyser les données de l'ECAD, il est utile de décrire les deux principaux types d'estimations ponctuelles des caractéristiques de la population qui peuvent être produites à partir du fichier de microdonnées créé pour l'ECAD.

### **8.3.1 Estimations catégorielles**

Les estimations catégorielles sont des estimations du nombre ou du pourcentage de la population visée par l'enquête possédant certaines caractéristiques ou faisant partie d'une catégorie définie. Le nombre de personnes qui ont déjà bu un verre ou la proportion d'utilisateurs de stimulants utilisant des stimulants prescrits constituent des exemples de telles estimations. Une estimation du nombre de personnes possédant une certaine caractéristique peut aussi être désignée comme une estimation d'un agrégat.

#### Exemples de questions catégorielles :

Q : Avez-vous **déjà** bu un verre?

R : Oui / Non

Q : Au cours des 12 derniers mois, est-ce que tous les stimulants, que vous avez utilisés, avaient été **prescrits pour vous**?

R : Oui, ils avaient tous été prescrits / Certains avaient été prescrits et d'autres non / Non, aucun n'avait été prescrit

### **8.3.2 Estimations quantitatives**

Les estimations quantitatives sont des estimations de totaux ou de moyennes, de médianes et d'autres mesures d'une tendance centrale de quantités reposant sur certains ou sur tous les membres de la population visée par l'enquête. Elles comprennent aussi expressément des estimations de la forme  $\hat{X}/\hat{Y}$  où  $\hat{X}$  est une estimation de la quantité totale pour la population visée par l'enquête et  $\hat{Y}$  est une estimation du nombre de personnes dans la population visée par l'enquête qui contribuent à cette quantité totale.

Un exemple d'estimation quantitative est le nombre moyen de verres consommés, au cours des 7 derniers jours, par personne. Le numérateur  $\left(\hat{X}\right)$  est une estimation du



nombre total de verres consommés au cours des 7 derniers jours et son dénominateur  $(\hat{Y})$  est le nombre de personnes ayant déclaré avoir consommé au moins une fois un verre au cours des 7 derniers jours.

Exemples de questions quantitatives :

Q : Au cours des 7 derniers jours, combien de verres avez-vous bu à chaque jour? (questions pour chacun des 7 derniers jours)

R : |\_|\_| verres

Q : À quel âge avez-vous essayé pour la première fois des amphétamines ou de la méthamphétamine?

R : |\_|\_|\_|ans

### 8.3.3 *Totalisation d'estimations catégorielles*

On peut obtenir des estimations du nombre de gens possédant une certaine caractéristique à partir du fichier de microdonnées en additionnant les poids finaux de tous les enregistrements possédant la ou les caractéristique(s) qui nous intéresse(nt). On obtient les proportions et les rapports de la forme  $\hat{X} / \hat{Y}$  en :

- additionnant les poids finaux des enregistrements présentant la caractéristique qui nous intéresse pour le numérateur  $(\hat{X})$ ,
- additionnant les poids finaux de tous les enregistrements pour le dénominateur  $(\hat{Y})$ , puis en
- divisant l'estimation a) par celle de b)  $(\hat{X} / \hat{Y})$ .

### 8.3.4 *Totalisation d'estimations quantitatives*

On peut obtenir des estimations de quantités à partir du fichier de microdonnées en multipliant la valeur de la variable qui nous intéresse par le poids final de chaque enregistrement, puis en additionnant cette quantité pour tous les enregistrements qui nous intéressent. Pour obtenir, par exemple, une estimation du nombre total de verres bus au cours des 7 derniers jours, multipliez la valeur déclarée à la question ALC\_70 (nombre de verres bus à chaque jour) par le poids final de l'enregistrement, puis additionnez cette valeur pour tous les enregistrements où la variable ALC\_70 < 96 (tous les répondants qui ont donné une réponse à ce champ).

Pour obtenir une moyenne pondérée de la forme  $\hat{X} / \hat{Y}$ , le numérateur  $(\hat{X})$  est calculé comme une estimation quantitative et le dénominateur  $(\hat{Y})$  est calculé comme une estimation catégorielle. Pour estimer, par exemple, le nombre moyen de verres bus au cours des 7 derniers jours,

- estimer le nombre total de verres bus au cours des 7 derniers jours,  $(\hat{X})$  tel qu'il est décrit ci-dessus,
- estimer le nombre de personnes  $(\hat{Y})$  incluses dans cette catégorie en additionnant les poids finaux de tous les enregistrements où la variable ALC\_70 < 96, puis
- diviser l'estimation a) par l'estimation b)  $(\hat{X} / \hat{Y})$ .

## 8.4 Lignes directrices pour l'analyse statistique

L'ECAD est basée sur un plan d'échantillonnage complexe, avec stratification, plusieurs étapes de sélection ainsi que des probabilités inégales de sélection des répondants. L'utilisation de données provenant d'enquêtes aussi complexes pose problèmes aux analystes car la conception de l'enquête et les probabilités de sélection affectent les procédures d'estimation et de calcul de la variance qui doivent être utilisées. Les poids de l'enquête doivent être utilisés lorsque des estimations sont produites ou des analyses effectuées.

Bien que de nombreuses procédures d'analyse figurant dans les logiciels statistiques permettent d'utiliser des poids, la signification ou la définition du poids dans ces procédures peut différer de celle qui est appropriée dans un cadre d'enquête par sondage sans utilisation des poids bootstrap. Dans de nombreux cas, les estimations produites par les logiciels sont correctes, mais si les variances ne sont pas basées sur les poids bootstrap, alors les variances calculées sont mauvaises.

Pour des analyses techniques plus complexes (par exemple la régression linéaire, la régression logistique et l'analyse de variance), il existe une méthode qui peut rendre les variances calculées par les ensembles standards plus significatifs (si les poids bootstrap ne sont pas utilisés), en incorporant les probabilités inégales de sélection. La méthode modifie les poids de manière à obtenir un poids moyen de 1.

Par exemple, supposons que l'analyse de tous les répondants masculins soit nécessaire. Les étapes pour modifier les poids sont les suivantes :

1. sélectionnez dans le fichier tous les répondants qui ont déclaré SEXE = homme ;
2. calculer le poids moyen pour ces enregistrements en additionnant les poids individuels originaux du fichier de microdonnées pour ces enregistrements et en divisant ensuite par le nombre de répondants qui ont déclaré SEXE = homme ;
3. pour chacun de ces répondants, calculer un poids MODIFIÉ égal au poids personne original divisé par le poids MOYEN ;
4. effectuer l'analyse pour ces répondants en utilisant le poids MODIFIÉ.

Toutefois, comme la stratification du plan de sondage n'est toujours pas prise en compte, les estimations de la variance ainsi calculées risquent d'être sous-estimées.

Dans la mesure du possible, les utilisateurs doivent utiliser les poids bootstrap dans leurs analyses afin d'estimer correctement les variances. Si l'utilisateur se sert d'un logiciel statistique qui permet l'analyse avec les poids bootstrap, il doit appliquer les poids bootstrap et non la méthode pour modifier les poids. Pour plus de détails sur l'utilisation des poids bootstrap dans le calcul de l'erreur d'échantillonnage utilisée dans les CV, les variances et les intervalles de confiance, veuillez consulter l'*Annexe A*.

Les paramètres pour le cycle 2019 de l'ECAD sont les suivants :

Pour le fichier maître :

- Fichier de données: CADS2019ECAD.txt
- Fichier de poids Bootstrap: CADS2019ECAD\_BSW.txt
- Variable d'identification variable: MASTERID
- Poids de l'enquête: WEIGHT
- Nombre de répliques bootstrap (B): 1000
- Poids des répliques : wrmp0001 to wrmp1000

Pour le FMGD :

- Fichier de données: CADS2019ECAD\_P.txt
- Fichier de poids Bootstrap: CADS2019ECAD\_P\_BSW.txt
- Variable d'identification variable: PUMFID
- Poids de l'enquête: WEIGHTP
- Nombre de répliques bootstrap (B): 1000
- Poids des répliques : wrpp0001 to wrpp1000

## 8.5 Lignes directrices de diffusion

### 8.5.1 Lignes directrices de diffusion basées sur la qualité

Avant de diffuser et/ou de publier des estimations de l'ECAD, les utilisateurs doivent tenir compte du niveau de qualité de l'estimation. La qualité des données est affectée par les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage, comme indiqué au *Chapitre 7.0*. Cette section traite de la qualité en termes d'erreurs d'échantillonnage. Il existe différentes façons de mesurer et de rapporter les erreurs d'échantillonnage. Statistique Canada considère comme une pratique exemplaire le fait de rapporter l'erreur d'échantillonnage d'une estimation par le biais de son intervalle de confiance à 95 %. L'intervalle de confiance doit être publié dans le même tableau que l'estimation. En plus des intervalles de confiance, les estimations sont classées dans l'une des trois catégories de qualité :

#### Catégorie A

Les estimations peuvent être publiées sans avertissement. Les utilisateurs de données peuvent utiliser l'intervalle de confiance de 95% pour décider si la qualité de l'estimation est suffisante.

#### Catégorie E – Qualité marginale

Les estimations et les intervalles de confiance sont jugés de qualité marginale. Les estimations et les intervalles de confiance doivent être signalés par la lettre E (ou un identificateur similaire) et être accompagnés d'un avertissement invitant à utiliser l'estimation avec prudence. Par exemple, « L'utilisateur est informé que les estimations et les intervalles de confiance marqués de la lettre E sont considérés comme étant de qualité marginale en raison de la forte variabilité d'échantillonnage, et doivent être utilisés avec prudence ».

#### Catégorie F – Qualité insuffisante

Les estimations et les intervalles de confiance sont jugés de mauvaise qualité. Les estimations présentent un niveau d'instabilité très élevé, ce qui les rend peu fiables et potentiellement trompeuses. Si les estimations sont publiées, elles doivent être accompagnées d'une clause de non-responsabilité. L'utilisateur doit reconnaître les avertissements donnés et s'engager à ne pas diffuser, présenter ou rapporter les estimations, directement ou indirectement, sans cette clause de non-responsabilité. Ils doivent être signalés par la lettre F (ou un autre identifiant similaire) et l'avertissement suivant doit accompagner les estimations et les intervalles de confiance :

« Veuillez noter que ces estimations et intervalles de confiance [marqués par la lettre F] ne répondent pas aux normes de qualité de Statistique Canada. Les conclusions basées sur ces données ne seront pas fiables et peuvent être invalides ».

Le tableau ci-dessous fournit les règles permettant d'attribuer à une estimation  $\hat{Y}$  et son intervalle de confiance à une catégorie de qualité (A, E ou F). Les règles sont principalement basées sur des comptes de l'échantillon.

**Tableau 2: Guide de diffusion**

| Type d'estimation | Catégorie A | Catégorie E | Catégorie F |
|-------------------|-------------|-------------|-------------|
|-------------------|-------------|-------------|-------------|

|   |  | <b>Qualité marginale</b> | <b>Qualité insuffisante</b>                    |
|---|--|--------------------------|--|
| Proportion                                    | $n \geq 163$                                   | Pas A et pas F           | $n < 82$                                       |
| Compte pondéré                                | $m \geq 163$                                   | Pas A et pas F           | $m < 82$                                       |
| Moyenne, $\hat{Y}$                            | $n \geq 163$ et $L \leq  \hat{Y} $             | Pas A et pas F           | $n < 82$ ou $L > 2 \hat{Y} $                   |
| Total, $\hat{Y}$                              | $m \geq 163$ et $L \leq  \hat{Y} $             | Pas A et pas F           | $m < 82$ ou $L > 2 \hat{Y} $                   |
| Différence, $\hat{Y} = \hat{Y}_1 - \hat{Y}_2$ | $\hat{Y}_1$ et $\hat{Y}_2$ sont de Catégorie A | Pas A et pas F           | $\hat{Y}_1$ ou $\hat{Y}_2$ sont de Catégorie F |

Notation:

$n$  : Taille de l'échantillon du domaine. Pour les proportions,  $n$  représente le compte non pondéré du nombre de répondants inclus dans le dénominateur de la proportion ; il n'y a pas d'exigence de taille d'échantillon pour le numérateur d'une proportion. Pour les moyennes,  $n$  représente le compte non pondéré du nombre de répondants qui contribuent au calcul de la moyenne (y compris les répondants avec des valeurs de zéro).

$m$  : Nombre non pondéré du nombre de répondants dont les valeurs sont différentes de zéro et qui contribuent à l'estimation

$L$  : longueur de l'intervalle de confiance à 95 % de  $\hat{Y}$ . La longueur de l'intervalle de confiance est utilisée pour les variables quantitatives telles que le revenu (par opposition aux variables dichotomiques ou catégorielles).

$|\cdot|$  : valeur absolue

Les règles du tableau 2 dépendent du type d'estimation. Les proportions et les comptes pondérés sont des estimations basées sur des variables dichotomiques ou catégorielles. Un exemple de compte pondéré est le nombre estimé de verres consommés. En revanche, les règles relatives aux moyennes et aux totaux s'appliquent aux variables quantitatives, telles que le revenu. Les estimations de la différence entre deux variables comprennent les estimations de la variation entre deux cycles d'enquête et les estimations de la différence entre deux domaines.

En plus des règles spécifiées par le tableau 2, deux conditions indiquent qu'un intervalle de confiance est de mauvaise qualité. La qualité d'une estimation et son intervalle de confiance doivent être classés comme étant de mauvaise qualité si l'une des deux conditions suivantes est vraie :

- La longueur de l'intervalle de confiance à 95 % est égale à zéro, c'est-à-dire  $L=0$  (sauf si l'estimation est basée sur un recensement plutôt que sur un échantillon, ou si l'estimation correspond à un total de contrôle d'étalonnage ; voir le *Chapitre 9* pour plus d'informations sur l'étalonnage).
- La limite inférieure ou supérieure de l'intervalle de confiance de 95% n'est pas une valeur plausible pour l'estimation. Cela indique que les hypothèses relatives à la distribution de l'estimation ne sont pas respectées. Par exemple, la limite inférieure de l'estimation du nombre de verres consommés ne doit pas être négative.

### **8.5.2 Lignes directrices de diffusion basées sur la confidentialité**

La section 8.5.1 établissait les lignes directrices de diffusion basées sur la qualité. Un autre aspect à considérer pour déterminer quelles estimations peuvent être diffusées est la confidentialité. Afin de s'assurer que l'identité des répondants est protégée, un minimum de 5 répondants doit contribuer à chaque estimation diffusée. Par exemple, ceci voudrait dire que le compte non pondéré de répondants inclus dans le numérateur d'une proportion est d'au moins 5.

## 9.0 Pondération

Dans le fichier de microdonnées, une série finale de poids individuels ont été attribués à chaque enregistrement pour indiquer le nombre que chaque personne échantillonnée représente. Des poids provisoires pour les ménages ont dû être calculés au préalable afin de calculer ces poids individuels.

La pondération pour les fichiers de l'ECAD comprend plusieurs étapes, la première consistant à calculer le poids des ménages :

- 1) calcul du poids initial : chaque ménage sélectionné représente plusieurs autres ménages dans sa strate;
- 2) élimination des enregistrements hors du champ de l'enquête;
- 3) ajustement pour les ménages non répondants (questions clés manquantes);
- 4) ajustement pour rendre les estimations des ménages cohérentes avec les totaux connus pour la province obtenus à partir de projections démographiques régulièrement produites par Statistique Canada.

Le calcul des poids des personnes commence avec les poids des ménages de l'étape 4 :

- 5) calcul du poids pour la sélection de la personne dans le ménage, basé sur la sélection selon l'âge;
- 6) ajustement pour rendre les estimations de la population cohérentes avec les totaux connus par province, groupe d'âge et sexe, ainsi que par Région Métropolitaine de Recensement (RMR), tirés des projections démographiques.

Toutes les étapes de la procédure de pondération sont décrites ci-dessous.

### 9.1 Procédures de pondération

#### 1. Calcul du poids initial

Un poids de base,  $W_1$ , a été attribué à chaque unité de l'échantillon, égal à l'inverse de sa probabilité de sélection dans chaque province.

$$W_{1,i} = \left( \frac{\text{Nombre d'unités pouvant être échantillonnées sur la base de sondage}}{\text{Nombre d'unités échantillonnées}} \right)$$

Il y avait 22 116 unités échantillonnées avec des poids attribués.

#### 2. Élimination des cas hors du champ de l'enquête

Les unités hors du champ de l'enquête, telles que celles dont l'adresse correspond à une entreprise, institution, logement saisonnier ou logement collectif ont été éliminées. 1 261 unités ont été identifiées comme étant hors du champ de l'enquête.

S'il y a des unités hors du champ de l'enquête,

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

Sinon,

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

#### 3. Ajustement pour les ménages non-répondants

Si aucun répondant n'a participé à l'enquête, par exemple à cause d'un refus de répondre, ou si la personne sélectionnée via la sélection du répondant selon l'âge n'a pas répondu aux questions

servant à la pondération (âge, taille du ménage, composition du ménage et à au moins deux questions portant sur la consommation d'alcool), alors le ménage (et la personne) a été considéré comme non-répondant. Il y a 10 562 ménages qui ont été considérés comme non-répondants. Plusieurs variables disponibles pour toutes les unités de l'échantillon ont été analysées pour être incluses dans le modèle de non-réponse, et celles qui avaient le plus grand pouvoir de prédiction ont été conservées. Les poids des 10 293 ménages répondants, dans le champ de l'enquête, ont donc été ajustés par revenu du ménage\*type de ménage\*présence d'un téléphone ou non. Le type de ménage représente la composition du ménage à l'intérieur du logement : personne vivant seule, couple sans enfants, couple avec enfants, un adulte avec enfant, etc.

$$W_{3,i} = W_{2,i} * \left( \frac{\sum W_2 \text{ pour les ménages répondants} + \sum W_2 \text{ pour les ménages non - répondants}}{\sum W_2 \text{ pour les ménages répondants}} \right)$$

#### 4. Ajustement pour totaux externes connus des ménages

Les poids des ménages pour chaque enregistrement ont été ajustés par province et taille du ménage pour s'assurer que les estimations des ménages étaient cohérentes avec des totaux externes connus des ménages. Ceci correspond au poids final des ménages. Le facteur d'ajustement par province\*taille du ménage a été défini comme suit :

$$W_{4,i} = W_{3,i} * \left( \frac{\text{Totaux externes connus de ménages}}{\sum W_3 \text{ pour les ménages répondants}} \right)$$

#### 5. Calcul du poids des personnes sélectionnées

Un poids a été attribué à toutes les personnes répondantes de l'enquête. Le poids initial de chaque personne est égal au poids final de son ménage, multiplié par l'inverse de sa probabilité d'avoir été sélectionné dans son ménage, selon la sélection basée sur l'âge :

$$W_{5,i} = W_{4,i} * \left( \frac{1}{\text{probabilité de sélection}} \right)$$

#### 6. Ajustement pour totaux externes

Un ajustement a été effectué sur les poids des personnes pour s'assurer que les estimations de la population étaient cohérentes avec les totaux externes de population pour les personnes de 15 ans et plus. Cette méthode est connue sous le nom de post-stratification. Les totaux externes suivants ont été utilisés :

- 1) totaux de population par province\*sexe\*groupe d'âge. Les groupes d'âge suivants ont été utilisés : 15 à 24 ans, 25 à 34 ans, 35 à 44 ans, 45 à 54 ans, 55 à 64 ans, 65 ans et plus;
- 2) totaux de population par RMR.

Les poids obtenus pour les personnes après cette étape sont considérés comme finaux au niveau de la personne et sont ceux qui figurent dans les fichiers de microdonnées.

## **10.0 Autre documentation**

### **Questionnaires :**

- Français : ECAD2019\_Questionnaire\_F.pdf
- Anglais : CADS2019\_Questionnaire\_E.pdf

### **Dictionnaires de données :**

- Fichiers de microdonnées à grande diffusion (FMGD)
  - Français : ECAD2019\_FMGD\_LvCd.pdf
  - Anglais : CADS2019\_PUMF\_Cdbk.pdf
- Fichiers maîtres
  - Français : ECAD2019\_MAIRE\_LvCd.pdf
  - Anglais : CADS2019\_MASTER\_Cdbk.pdf



## ***Annexe A – Estimation de la variance et construction des intervalles de confiance***

Afin de mesurer l'erreur d'échantillonnage des estimations, il faut calculer les estimations de la variance et il faut construire les intervalles de confiance. L'ECAD utilise une méthode complexe pour concevoir le plan d'échantillonnage et l'estimation, 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, nommés WRMP1-WRMP1000. Essentiellement, la variance est estimée en calculant la valeur de l'estimation souhaitée au moyen de chacun des ensembles de poids bootstrap, puis la variabilité entre ces estimations bootstrap est ensuite mesurée.

### **Progiciels statistiques pour estimer la variance**

Pour l'ECAD, il est nécessaire d'utiliser les poids bootstrap pour calculer des estimations de la variance exactes. Un certain nombre de programmes ou de logiciels 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 logiciels d'analyse standard ou plus anciens, dont SPSS, SAS avant la version 9.2, ne comportent pas de procédure intégrée d'estimation de la variance à partir de poids bootstrap lorsqu'on utilise des données fondées sur un plan de sondage complexe comme celui de l'ECAD. Ces logiciels ne devraient 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 les variances à partir des poids bootstrap ainsi que d'autres types de poids de rééchantillonnage comme le Jackknife et les 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 les poids de rééchantillonnage ne sont offerts que dans les versions 9.3 et plus récentes de SAS.

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).

### **Intervalles de confiance**

La méthode la plus couramment utilisée pour construire des intervalles de confiance à 95 % est l'intervalle de Wald, qui est de la forme  $\hat{y} \pm 1,96\sqrt{\hat{v}\hat{a}r(\hat{y})}$  pour une estimation  $\hat{y}$  avec l'estimation de variance  $\hat{v}\hat{a}r(\hat{y})$ . Les intervalles de Wald sont fondés sur l'hypothèse selon laquelle la distribution d'échantillonnage de  $\hat{y}$  est approximativement normale. Pour les proportions, l'hypothèse de normalité ne tient généralement plus pour les échantillons de petite taille et pour des proportions proches de zéro ou de un. Trois méthodes de rechange 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; 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 d'autres méthodes de construction des intervalles de confiance sont précisées pour des proportions dans SAS et SUDAAN.

1. SAS, intervalles de confiance de Wilson modifiés  
PROC SURVEYFREQ  
DATA=.... VARMETHOD=BRR;  
WEIGHT WEIGHT;  
REPWEIGHTS WRMP1-WRMP1000;  
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 WEIGHT;  
REPWGT WRMP1-WRMP1000;  
TABLES ...;

## Références

- Korn, E.L. et B.I. Graubard. 1998. « Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data », *Survey Methodology*, vol. 24, p. 193 à 201.
- Liu, Y.K. et P.S. Kott. 2009. « Evaluating Alternative One-Sided Coverage Intervals for a Proportion », *Journal of Official Statistics*, vol. 25, n° 4, p. 569 à 588.
- Phillips, O. 2004. « Comment utiliser les poids bootstrap avec WesVar et SUDAAN », (n° de catalogue 12-002-X20040027032) Bulletin technique et d'information des centres de données de recherche, index chronologique, automne 2004, vol.1, n° 2, Statistique Canada, n° de catalogue 12-002-XIE.