Nº au catalogue 89-657-X2025002
ISSN 2371-5014
ISBN 978-0-660-75073-6

Série thématique sur l'ethnicité, la langue et l'immigration

Enquête sur la population de langue officielle en situation minoritaire : guide de l'utilisateur, 2022 (révisions de 2025)

Date de diffusion : le 14 novembre 2025

Statistique Canada
Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par le ministre responsable de Statistique Canada

Une version HTML est aussi disponible.

This publication is also available in English.

Table des matières

Enquête sur la population de langue officielle en situation minoritaire : guide de l'utilisateur, 2022 (révisions de 2025)

1.0 Description de l'enquête

L'Enquête sur la population de langue officielle en situation minoritaire (EPLOSM) a été menée par Statistique Canada en 2022 avec la collaboration et le soutien de Patrimoine canadien. Il s'agit d'une enquête postcensitaire réalisée auprès de la population de langue anglaise du Québec et de la population de langue française ailleurs au Canada. Les questions ont été conçues pour évaluer l'évolution de la situation des populations de langue officielle en situation minoritaire depuis 2006, année où une enquête semblable (l'Enquête sur la vitalité des minorités de langue officielle [EVMLO]) a été menée par Statistique Canada, et pour fournir des renseignements sur les nouveaux enjeux concernant ces populations minoritaires. L'enquête comprend deux grands échantillons : l'un pour les adultes appartenant à la population de langue officielle en situation minoritaire et l'autre pour les enfants dont au moins un parent appartient à une population de langue officielle en situation minoritaire ou qui sont eux-mêmes admissibles à l'instruction dans la langue officielle minoritaire.

Les résultats de l'enquête ont été produits pour aider les gouvernements fédéral, provinciaux et territoriaux et les administrations publiques municipales et les organismes communautaires à fournir des services, des programmes et des initiatives pour les minorités de langue officielle. Les données sont également utiles aux chercheurs, aux chercheuses et aux autres intervenants qui s'intéressent à divers aspects du français au Canada hors du Québec et de l'anglais au Québec.

Toute question concernant l'ensemble de données ou son utilisation doit être adressée à :

Statistique Canada
Centre de démographie, Statistique linguistique
Services à la clientèle
Courriel : statcan.languagestats-statlinguistique.statcan@statcan.gc.ca

2.0 Concepts et définitions

2.1 Concepts et définitions de l'Enquête sur la population de langue officielle en situation minoritaire

Dans cette enquête, la population de langue officielle en situation minoritaire désigne la population de langue anglaise au Québec et la population de langue française au Canada hors du Québec.

Dans ce contexte, la sélection de l'échantillon exigeait le respect de critères précis. Ces critères, présentés en détail à la section 3.1 (Population cible et population observée), ont été déterminés en consultation avec le comité consultatif externe d'experts mis en place aux fins de l'enquête. Les critères ont été sélectionnés pour répondre aux paramètres suivants : permettre des comparaisons avec l'enquête de 2006; être suffisamment généraux pour que les utilisateurs puissent préciser des sous-populations en fonction de critères linguistiques précis; et pour les enfants, tenir compte des nouveaux renseignements sur la langue d'instruction. Les critères ne visaient pas à créer une définition officielle de la population de langue officielle en situation minoritaire.

Les concepts linguistiques (langue maternelle, langue parlée le plus souvent à la maison, connaissance des langues officielles, admissibilité à l'instruction dans la langue officielle minoritaire) sur lesquels reposent les critères sont tous des concepts tirés du recensement. Veuillez consulter les guides de référence du recensement pour obtenir les définitions et les notes sur la qualité relatives à ces concepts :

- Guide de référence sur les langues, Recensement de la population, 2021;
- Guide de référence sur l'instruction dans la langue officielle minoritaire, Recensement de la population, 2021.

L'EPLOSM a été conçue pour recueillir des renseignements approfondis sur divers aspects des populations de langue officielle en situation minoritaire afin de compléter l'information provenant des sources de données existantes, dont le recensement. En plus de renseignements sociodémographiques et des liens entre les membres des ménages, l'enquête couvre les sujets principaux suivants :

- compétences linguistiques;
    - la capacité du répondant à comprendre, à parler, à lire et à écrire en français ou en anglais, ainsi que l'hésitation à utiliser la langue officielle minoritaire et l'insécurité et la discrimination linguistiques perçues;
- scolarité;
    - les diplômes obtenus et la fréquentation scolaire, ce qui comprend la langue d'instruction, les raisons du choix de la langue d'instruction et les intentions des parents à l'égard de la langue d'instruction de leurs enfants;
- services à la petite enfance;
    - comprend la langue des services de garde et les raisons du choix de la langue des services de garde;
- trajectoire linguistique de l'enfance à l'âge adulte et dynamique linguistique familiale de l'enfant;
    - les langues parlées dans la petite enfance et à l'âge de 15 ans, et des renseignements détaillés sur les langues parlées à la maison et les langues parlées avec des amis;
- sentiment d'appartenance et vitalité perçue;
    - le sentiment d'appartenance au groupe linguistique, l'importance du français ou de l'anglais et de leur promotion pour les répondants, et la présence locale perçue du français ou de l'anglais lorsque le répondant était âgé de 15 ans et au moment de remplir le questionnaire et 10 ans auparavant;
- participation communautaire;
    - la participation à des organismes, des associations et des réseaux au sein de la collectivité;
- utilisation des langues dans l'espace public;
    - les langues parlées à l'extérieur de la maison et du travail, avec des personnes autres que des membres de la famille ou des amis, dans des endroits comme les épiceries, les restaurants et les hôpitaux;
- services gouvernementaux;
    - la langue de communication avec les différents ordres de gouvernement (provincial ou territorial, fédéral) et les administrations publiques municipales;
- contacts avec le système de justice;
    - la langue des interactions avec le système de justice;
- services de santé;
    - la langue de service dans le domaine de la santé;
- services aux immigrants;
    - la langue des services obtenus lors de l'immigration;
- art, culture et médias;
    - la langue employée pour utiliser des médias et Internet, participer à des sports organisés, assister à des activités artistiques, etc.;
- mobilité géographique;
    - le lieu de naissance, de résidence à l'âge de 15 ans, de résidence au moment de l'enquête, les raisons et les intentions de la mobilité géographique;
- participation au marché du travail;
    - des renseignements détaillés sur les langues utilisées au travail, les obstacles perçus au travail liés aux langues, et les services de recherche d'emploi.

Des questions liées à la pandémie de COVID-19 ont été ajoutées au questionnaire. Elles visent principalement à évaluer la mesure dans laquelle la pandémie a influé sur les résultats de l'enquête et pourrait avoir contribué à des changements depuis l'EVMLO de 2006, particulièrement dans les domaines des services en français et en anglais et de la participation aux activités.

Les questions de l'EPLOSM renvoient parfois aux concepts de « communauté » et de « municipalité ». Ces concepts sont définis dans le questionnaire ainsi :

- « Le terme « communauté » désigne notamment les endroits de votre quartier que vous fréquentez régulièrement et qui se trouvent à distance de marche ou d'un court trajet en voiture, comme les épiceries, les écoles, les pharmacies ou les centres de conditionnement physique. »
- « Le terme « municipalité » fait référence à la ville, la municipalité ou la localité de résidence aux termes des lois provinciales ou territoriales. »

Le dictionnaire de données ainsi que la version complète du guide de l'utilisateur seront accessibles aux utilisateurs des microdonnées, lorsque les fichiers de microdonnées seront disponibles.

Pour la description détaillée des concepts associés aux questions de l'EPLOSM, veuillez consulter les documents suivants :

- Enquêtes et programmes statistiques - Enquête sur la population de langue officielle en situation minoritaire (EPLOSM)
- Questionnaire(s) et guide(s) de déclaration – Enquête sur la population de langue officielle en situation minoritaire

2.2 Élaboration du contenu

Le contenu de l'EPLOSM de 2022 a été fondé sur la version précédente de cette enquête postcensitaire, l'EVMLO, qui a été réalisée en 2006. La comparabilité historique a été désignée comme une priorité. Bien qu'une majeure partie du contenu soit répété, des mises à jour au questionnaire ont été effectuées à la suite de consultations avec Patrimoine canadien et les principaux utilisateurs de données, c'est-à-dire des partenaires fédéraux et communautaires et des comités consultatifs externes.

Ces mises à jour visaient à répondre aux nouveaux besoins en matière de données et de mieux comprendre les communautés de langue officielle en situation minoritaire au Canada. Quelques exemples comprennent un module sur l'expérience des immigrants en matière de services d'aide à l'intégration, et de nouvelles questions portant sur des sujets comme l'hésitation à utiliser la langue officielle minoritaire et l'insécurité et la discrimination linguistiques. De plus, un nouveau module sur les répercussions de la pandémie de COVID-19 a été créé pour évaluer les effets de la pandémie sur les activités et l'accès aux services dans la langue officielle minoritaire.

En outre, le questionnaire de l'enquête de 2006 a été examiné dans le but de réduire le plus possible le fardeau de réponse. Environ 80 % des questions et des concepts du questionnaire de 2006 sont inclus dans l'enquête de 2022. En ce qui a trait aux questions qui ont été retirées, une partie du contenu peut être dérivé, partiellement ou entièrement, au moyen du couplage d'enregistrements avec les données du recensement, ce qui réduit le fardeau de réponse.

Voici quelques exemples de questions retirées :

- l'instruction des parents dans la langue officielle minoritaire, le type de programme en français fréquenté dont les programmes d'immersion et le nombre d'années au sein d'un programme dans la langue officielle minoritaire;
- si l'enfant vit avec ses deux parents et, dans le cas contraire, les raisons, la durée et la fréquence des séjours chez chaque parent;
- année scolaire de l'enfant, date d'inscription, mode de transport pour se rendre à l'école et en revenir, durée du trajet, nombre d'années dans un programme d'immersion et raisons du changement de programme;
- le revenu personnel et le revenu du ménage.

Deux rondes des tests qualitatifs ont été menés en août 2020 et en mars 2021 auprès de répondants qui respectaient les critères de participation à l'enquête. L'objectif était de s'assurer que toutes les questions, en particulier les questions nouvelles et modifiées qui visent à répondre aux nouveaux besoins en matière de données, étaient faciles à comprendre et pouvaient fournir des données de qualité. Les essais ont été organisés par des spécialistes de la conception de questionnaires. Les essais qualitatifs se sont déroulés sur une période de deux semaines chacun et environ 30 interviews ont eu lieu en français et en anglais. Chaque interview durait une heure environ. Dans l'ensemble, l'enquête a été bien accueillie par les répondants; les questions étaient bien comprises et, à part quelques suggestions sur différents points de l'enquête, les commentaires étaient positifs.

La plupart des modules ont fait l'objet de modifications mineures, mais certains ont nécessité des changements plus importants à la suite des essais. Une modification importante concernait le module de la trajectoire linguistique, qui était à l'origine le dernier module du questionnaire. Les répondants estimaient que le module était redondant lorsqu'il était placé à la fin du questionnaire. Ainsi, il a été déplacé pour figurer plus tôt dans le questionnaire. Des changements ont aussi été apportés au module lui-même. Les répondants ont aussi considéré que le module sur la mobilité géographique était trop compliqué et qu'il était difficile d'y répondre. C'est pourquoi d'importantes modifications ont été apportées pour le simplifier et le raccourcir. Enfin, la section sur la scolarité a aussi été modifiée pour faciliter la compréhension.

3.0 Méthodologie de l'enquête

La méthodologie d'enquête et l'analyse de données d'enquêtes complexes sont des sujets vastes et techniques, chacun le sujet d'ouvrages entiers (qui existent déjà). Des aperçus courts, génériques et non techniques des deux sujets, qui se veulent des introductions aux descriptions plus exhaustives que l'on retrouve dans la littérature, se trouvent dans les annexes C et E. L'annexe C décrit la méthodologie générale, alors que les sections suivantes du présent guide fournissent des renseignements supplémentaires pertinents pour l'EPLOSM.

Les principaux points à retenir de l'annexe C sont les suivants :

- L'inexactitude d'une estimation comporte deux composantes principales : le biais et la variance.
- Contrairement à la variance, le biais est difficile, voire impossible, à évaluer en pratique.
- Pour l'EPLOSM, comme pour les autres enquêtes de Statistique Canada, des méthodes éprouvées et des pratiques exemplaires ont été mises en œuvre afin de réduire les sources de biais au sein des données.
- Le poids d'enquête s'emploie avec les données afin de réduire le biais des estimations produites.
- Les poids Bootstrap doivent être utilisés afin d'évaluer la variance de ces estimations.
- La variance associée à une estimation est communiquée sous la forme d'un intervalle de confiance valide.
- Plus un intervalle de confiance valide est étroit, plus précise est vraisemblablement l'estimation.

3.1 Population cible et population observée

Depuis 1969, la *Loi sur les langues officielles* reconnaît le français et l'anglais comme les **deux langues officielles** du pays. Pour les personnes vivant au Québec, l'anglais est la langue officielle minoritaire, tandis que le français est la langue officielle minoritaire des personnes vivant dans les autres provinces ou territoires[1].

Dans le contexte de l'EPLOSM, une **personne est considérée comme faisant partie de la population de langue officielle en situation minoritaire** si elle répond à l'un des critères suivants :

1. sa langue maternelle comprend la langue officielle minoritaire;
2. sa langue maternelle ne comprend ni l'anglais ni le français, et la seule langue officielle qu'elle connaît est la langue officielle minoritaire;
3. sa langue maternelle ne comprend ni l'anglais ni le français, elle connaît les deux langues officielles, et ne parle pas la langue officielle majoritaire le plus souvent à la maison.

1. Dans le cadre de cette enquête, les langues officielles minoritaires du Canada sont l'anglais au Québec et le français au Canada hors Québec, bien qu'à l'échelle des provinces et des territoires, les langues ayant un statut officiel puissent différer.

Deux segments principaux de la **population sont couverts** par l'EPLOSM : les adultes et les enfants.

Une personne est considérée comme faisant partie de **la composante des adultes de l'EPLOSM** si elle :

1. avait 18 ans et plus le 16 mai 2022 (premier jour de la collecte de données de l'EPLOSM);
2. réside dans l'une des 10 provinces ou l'une des trois capitales territoriales, et ne vit pas dans un logement collectif, une réserve des premières nations ou une communauté inuite dans le nord du Québec;
3. est canadienne (les résidents non permanents sont exclus);
4. fait partie de la population de langue officielle en situation minoritaire.

La **composante des enfants de l'EPLOSM** se compose des deux parties ci-dessous.

Enfants vivant avec un parent qui fait partie de la population de langue officielle en situation minoritaire

Dans cette composante, les enfants doivent respecter les critères suivants :

1. être âgés de moins de 18 ans le 16 mai 2022;
2. résider dans l'une des 10 provinces ou l'une des trois capitales territoriales, et ne pas vivre dans un logement collectif, une réserve des premières nations ou une communauté inuite du nord du Québec;
3. ne pas être parent;
4. avoir au moins un parent qui répond aux exigences de la composante des adultes, sous réserve que le parent soit une personne âgée de 15 ans ou plus.

Il importe de souligner qu'un enfant est considéré comme faisant partie de la population cible de l'enquête de l'EPLOSM selon le profil linguistique de ses parents et non du sien. Par conséquent, il se peut que certains enfants de l'échantillon ne parlent pas ou ne comprennent pas la langue officielle minoritaire de leur province ou territoire de résidence.

Enfants admissibles à l'instruction dans la langue officielle minoritaire

Dans cette partie, les enfants doivent respecter les critères suivants :

1. avoir moins de 18 ans le 16 mai 2022;
2. résider dans l'une des 10 provinces ou l'une des trois capitales territoriales et ne pas vivre dans un logement collectif, une réserve des premières nations ou une communauté inuite du nord du Québec;
3. ne pas être parent.

De plus, ils doivent aussi répondre à l'un des critères suivants :

4a. vivre à l'extérieur du Québec dans la même famille de recensement qu'une personne âgée de 15 ans ou plus qui a le français comme langue maternelle;

4b. vivre dans la même famille de recensement qu'une personne âgée de 15 ans ou plus qui est ou a été instruit à l'école primaire dans la langue officielle minoritaire de sa province ou de son territoire;

5. être ou avoir été instruit dans la langue officielle minoritaire de leur province ou territoire au niveau primaire ou secondaire;
6. vivre avec un frère ou une sœur au sein de la même famille de recensement qui est ou a été instruit dans la langue officielle minoritaire de sa province ou de son territoire au niveau primaire ou secondaire.

Le concept de parentalité est utilisé ici au sens large. Au-delà de ceux qui ont déclaré être des parents dans le recensement, l'EPLOSM permet aux personnes canadiennes âgées de 15 ans et plus qui vivent dans la même famille de recensement d'être considérées comme parent. Ainsi, la définition des enfants admissibles à l'instruction dans la langue officielle minoritaire est délibérément plus large que celle du Recensement de 2021. Cela s'explique principalement par le fait que les spécifications pour la définition du Recensement de 2021 n'étaient pas terminées au moment où l'échantillon de l'EPLOSM a été tiré. Une définition plus large permettait de couvrir divers scénarios

au moyen des données de l'EPLOSM, dont la définition adoptée par la suite dans le recensement[2], ainsi que d'autres définitions d'intérêt pour les utilisateurs. La différence entre les populations de la définition la plus large et la plus stricte est estimée à 1,7 %.

En revanche, il se peut que certains enfants admissibles à l'instruction dans la langue officielle minoritaire en vertu de l'article 23 de la *Charte canadienne des droits et libertés* et de sa jurisprudence ne soient pas classés comme tels dans le recensement et que, par conséquent, ils ne soient pas inclus dans l'échantillon de l'EPLOSM. Bien qu'en vertu de la loi, l'admissibilité d'un enfant à l'instruction dans la langue officielle minoritaire dépend, entres autres, des caractéristiques de ses parents et de ses frères et sœurs, les données du recensement établissent les relations familiales seulement pour les personnes vivant dans le même ménage. Pour en savoir plus sur cette limite, consultez l'Étude sur la sous-classification des enfants admissibles à l'instruction dans la langue officielle minoritaire au Recensement de 2021.

Additionnellement, en raison de contraintes opérationnelles dans les territoires, seules les personnes résidant dans les capitales étaient admissibles à l'enquête. D'après les données du recensement, environ 77 % des francophones des territoires résident dans les capitales. De plus, conformément aux exigences indiquées ci-dessus, l'EPLOSM exclut les personnes non canadiennes et celles qui vivent dans un logement collectif, dans une réserve des premières nations ou dans une communauté inuite du nord du Québec.

3.2 Plan de sondage

La base de sondage de l'EPLOSM est dérivée à partir des données recueillies dans le cadre du Recensement de la population de 2021, tenu le 11 mai 2021, ce qui fait de l'EPLOSM une enquête postcensitaire.

Le recensement de 2021 utilise deux questionnaires[3] pour recueillir des données auprès des personnes vivant dans un logement : le questionnaire abrégé (formulaire 2A) et le questionnaire détaillé (formulaire 2A-L ou formulaire 2A-R selon l'emplacement géographique du logement). Le formulaire 2A-L est envoyé à environ 1 logement privé sur 4 dans la plupart des régions du Canada. Le questionnaire abrégé (formulaire 2A) comprend les questions démographiques de base du recensement (nom, sexe à la naissance et genre, date de naissance et âge, état matrimonial légal, union libre, relation à la Personne 1, connaissance des langues officielles, langues parlées à la maison, première langue apprise, langue d'instruction et expérience militaire canadienne). Le formulaire détaillé comprend aussi les questions de base ainsi que des questions sur l'activité sur le marché du travail, le revenu, la scolarité, la citoyenneté, le logement, les origines ethniques ou culturelles, la religion et l'identité autochtone, entre autres. Le formulaire 2A-R est semblable au formulaire 2A-L et cible tous les ménages dans les communautés des Premières Nations, les établissements métis, les régions inuites et d'autres régions éloignées[4].

Les citoyens canadiens qui vivent temporairement à l'étranger, les membres à temps plein des Forces armées canadiennes en poste à l'étranger et les visiteurs ou représentants de gouvernements étrangers sont exclus de la population cible du questionnaire détaillé du recensement. Sont également exclues les personnes vivant dans des logements collectifs (institutionnels ou non). Les logements collectifs comprennent les hôpitaux, les résidences pour personnes âgées, les établissements de soins pour bénéficiaires internes comme les foyers collectifs pour les personnes ayant une incapacité ou une dépendance, les refuges, les établissements correctionnels et de détention, les pensions et les maisons de chambres, les établissements religieux, les colonies huttérites, les établissements offrant des services d'hébergement temporaire[5] et d'autres établissements[6]. Pour en savoir plus sur les logements collectifs, consultez le Dictionnaire, Recensement de la population, 2021.

Bien que les questions du recensement sur les langues, soit celles qui sont d'un intérêt particulier pour l'EPLOSM, se trouvent dans les questionnaires abrégé et détaillé, l'échantillon de l'EPLOSM a été sélectionné parmi les personnes ayant répondu au questionnaire détaillé pour que les réponses de l'EPLOSM soient enrichies par les réponses aux questions additionnelles du formulaire détaillé. Exceptionnellement, quand le nombre de répondants

2. Pour en savoir plus sur les concepts connexes du recensement, veuillez consulter le Guide de référence sur l'instruction dans la langue officielle minoritaire, Recensement de la population, 2021.
3. Voir les questionnaires du Recensement de 2021.
4. Le présent document utilise l'expression « régions éloignées » pour désigner toutes ces régions afin d'alléger le texte.
5. Par exemple, un hôtel, un terrain de camping, un YMCA ou un YWCA, un manoir Ronald McDonald ou une auberge, s'il s'agit du lieu habituel de résidence du répondant le jour du recensement.
6. Comme une résidence pour étudiants, une base militaire, un camp de chantier ou un navire.

du questionnaire détaillé était insuffisant pour répondre aux exigences analytiques ciblées de l'EPLOSM, l'échantillonnage a été réalisé en puisant dans le bassin plus grand des répondants au questionnaire abrégé.

La base de sondage de l'EPLOSM a été stratifiée par région et groupe d'âge. Les strates ont également tenu compte de caractéristiques linguistiques de la manière suivante dans les segments adulte et enfant de la base de sondage :

- Pour les adultes, la stratification est fonction de leur groupe linguistique
- Pour les enfants, la stratification prend en compte leur admissibilité à l'instruction dans la langue officielle minoritaire ainsi que le profil linguistique de leurs parents.

Comme en 2006, les provinces du Québec, de l'Ontario et du Nouveau-Brunswick ont été divisées respectivement en six, cinq et trois régions, comme l'indique le tableau suivant. Dans toutes les autres provinces, la région correspond à la province. Pour les trois territoires, la capitale constitue la seule région de chaque territoire. Il s'agit de Whitehorse (AR) au Yukon, de Yellowknife (AR) aux Territoires du Nord-Ouest et d'Iqaluit (SDR) au Nunavut.

Tableau 1
Régions de l'EPLOSM et la géographie du recensement correspondante, pour le Nouveau-Brunswick, le Québec et l'Ontario

Province	Région métropolitaine de recensement (RMR), division de recensement (DR) ou subdivision de recensement (SDR)[1]	Région
Nouveau-Brunswick	1312 (Victoria), 1313 (Madawaska), 1314 (Restigouche), 1315 (Gloucester), 1309036 (Alnwick), 1309038 (Neguac)	Nord
	1307 (Westmorland), 1308 (Kent), 1309016 (Rogersville, Paroisse), 1309017 (Rogersville, Village), 1309001 (Hardwicke)	Sud-Est
	Toutes les autres qui ne figurent pas dans la liste ci-dessus	Reste
Québec	462 (Montréal)	Montréal
	2401 (Îles-de-la-Madeleine), 2402 (Le Rocher-Percé), 2403 (La Côte-de-Gaspé), 2404 (La Haute-Gaspésie), 2405 (Bonaventure), 2406 (Avignon), 2495 (La Haute-Côte-Nord), 2496 (Manicouagan), 2497 (Sept-Rivières–Caniapiscau), 2498 (Minganie–Le Golfe-du-Saint-Laurent), 2407 (La Matapédia), 2408 (La Matanie), 2409 (La Mitis), 2410 (Rimouski-Neigette), 2411 (Les Basques), 2412 (Rivière-du-Loup), 2413 (Témiscouata), 2414 (Kamouraska)	Est
	2480 (Papineau), 2481 (Gatineau), 2482 (Les Collines-de-l'Outaouais), 2483 (La Vallée-de-la-Gatineau), 2484 (Pontiac), 2485 (Témiscamingue), 2486 (Rouyn-Noranda), 2487 (Abitibi-Ouest), 2488 (Abitibi), 2489 (La Vallée-de-l'Or)	Ouest
	2430 (Le Granit), 2440 (Les Sources), 2441 (Le Haut-Saint-François), 2442 (Le Val-Saint-François), 2443 (Sherbrooke), 2444 (Coaticook), 2445 (Memphrémagog), 2446 (Brome-Missisquoi), 2447 (La Haute-Yamaska), 2448 (Acton), 2456 (Le Haut-Richelieu), 2468 (Les Jardins-de-Napierville), 2469 (Le Haut-Saint-Laurent)	Estrie et Sud
	2415 (Charlevoix-Est), 2416 (Charlevoix), 2420 (L'Île-d'Orléans), 2421 (La Côte-de-Beaupré), 2422 (La Jacques-Cartier), 2423 (Québec), 2434 (Portneuf), 2417 (L'Islet), 2418 (Montmagny), 2419 (Bellechasse), 2425 (Lévis), 2426 (La Nouvelle-Beauce), 2427 (Robert-Cliche), 2428 (Les Etchemins), 2429 (Beauce-Sartigan), 2431 (Les Appalaches), 2433 (Lotbinière), 2491 (Le Domaine-du-Roy), 2492 (Maria-Chapdelaine), 2493 (Lac-Saint-Jean-Est), 2494 (Le Saguenay-et-son-Fjord)	Québec et environs
	Toutes les autres qui ne figurent pas dans la liste ci-dessus	Reste
Ontario	3501 (Stormont, Dundas et Glengarry), 3502 (Prescott et Russell)	Sud-Est
	3506 (Ottawa)	Ottawa
	3548 (Nipissing), 3552 (Sudbury), 3553 (Greater Sudbury / Grand Sudbury), 3554 (Timiskaming), 3556 (Cochrane), 3557 (Algoma)	Nord-Est
	3520 (Toronto)	Toronto
	Toutes les autres qui ne figurent pas dans la liste ci-dessus	Reste

1. Pour en savoir plus, consultez 'Classification géographique type (CGT) 2021 - Volume I, La classification'.
Source : Classification géographique type (CGT) 2021 - Volume I, La classification.

Les groupes d'âge pour les provinces étaient de 1 à 4 ans, de 5 à 11 ans, de 12 à 17 ans, de 18 à 24 ans, de 25 à 44 ans, de 45 à 64 ans et de 65 ans et plus; les groupes d'âge pour les capitales des territoires étaient de 17 ans ou moins et de 18 ans ou plus.

Le groupe linguistique est formé de personnes qui font partie de la population de langue officielle en situation minoritaire, selon la définition donnée plus haut. Cependant, dans la région de Montréal, le groupe linguistique de base de la province de Québec a été affiné à des fins de stratification en y ajoutant les deux strates suivantes :

Les individus dont

- Langue maternelle est l'anglais et une autre langue non officielle, et dont la langue officielle connue est uniquement l'anglais; ou,
- Langue maternelle est l'anglais et une autre langue non officielle, et dont les deux langues officielles sont connues; ou,
- Langue maternelle comprend l'anglais, le français et une autre langue, et les deux langues officielles sont connues, et la langue parlée à la maison le plus souvent n'est pas le français; ou,
- La langue maternelle n'inclut aucune langue officielle et la langue officielle connue est uniquement l'anglais; ou,
- La langue maternelle n'inclut aucune langue officielle, et les deux langues officielles sont connues, et la langue parlée le plus souvent à la maison est l'anglais, mais pas le français.

Les individus dont

- La langue maternelle n'inclut aucune langue officielle, et les langues officielles connues sont le français et l'anglais, et la langue parlée le plus souvent à la maison est une langue tierce (autre que le français ou l'anglais); ou,
- La langue maternelle comprend les deux langues officielles ainsi qu'une langue tierce, et les deux langues officielles sont connues, et les langues parlées à la maison le plus souvent sont les deux langues officielles; ou,
- La langue maternelle comprend les deux langues officielles ainsi qu'une langue tierce, et les deux langues officielles sont connues, et la langue parlée à la maison le plus souvent est une langue tierce.

A noter qu'en 2006, pour des besoins analytiques spécifiques, on avait ajouté un échantillon distinct de personnes de langue maternelle tierce et ayant le français comme première langue officielle parlée dans la région de Montréal. En 2022 il n'y a pas eu de tel ajout.

Un échantillon aléatoire simple sans remise a été sélectionné dans chaque strate et indépendamment entre strates. Pour certaines strates, y compris celles de Terre-Neuve-et-Labrador et de l'Île-du-Prince-Édouard, la taille de l'échantillon ciblé dépassait le nombre de personnes disponibles ayant répondu au questionnaire détaillé. Dans ces cas, la stratégie d'échantillonnage a été ajustée pour inclure des personnes ayant répondu au questionnaire abrégé. Par conséquent, les données du questionnaire détaillé ne seront disponibles dans le fichier analytique que pour environ le quart des personnes sélectionnées dans ces strates.

Dans une même région, les expériences de la collectivité peuvent varier en fonction de la concentration de la population de langue officielle en situation minoritaire, mesurée par la proportion de personnes de la langue officielle minoritaire vivant dans chaque aire de diffusion (AD) du recensement. Le plan d'échantillonnage assure que dans chacune des strates l'échantillon représente bien toute la diversité des niveaux de concentration en empêchant l'abondance d'AD ayant des niveaux de concentration similaires.

3.3 Taille de l'échantillon

La taille d'échantillon pour chaque strate a été déterminée en ciblant un niveau d'exactitude comparable à celui de l'édition précédente de l'enquête, soit l'EVMLO de 2006. Ainsi, pour l'EPLOSM on a prévu d'estimer une proportion d'environ 8 % avec un coefficient de variation de 25 %, ou de manière équivalente, avec une marge d'erreur d'environ 4 points de pourcentage ou un intervalle de confiance d'une longueur d'environ 8 points de pourcentage.

Le tableau suivant présente les tailles d'échantillon par région, pour les composantes des adultes et des enfants de l'EPLOSM.

Tableau 2
Les tailles d'échantillon pour les composantes des adultes et des enfants et pour l'ensemble de l'enquête, selon la province, la capitale territoriale ou la région

Région	Adultes	Enfants	Région	Province ou capitale (territoire)
Terre-Neuve-et-Labrador	949	964	1 913	1 913
Île-du-Prince-Édouard	1 074	916	1 990	1 990
Nouvelle-Écosse	1 263	1 448	2 711	2 711
Nouveau-Brunswick, Nord	1 297	971	2 268	6 910
Nouveau-Brunswick, Sud-Est	1 293	1 006	2 299	
Nouveau-Brunswick, Reste	1 236	1 107	2 343	
Québec, Est	1 110	1 026	2 136	20 209
Québec, Estrie et Sud	1 281	1 343	2 624	
Québec, Montréal	3 942	3 119	7 061	
Québec, Ouest	1 296	1 354	2 650	
Québec, Québec et environs	1 262	1 691	2 953	
Québec, Reste	1 264	1 521	2 785	
Ontario, Nord-Est	1 297	1 163	2 460	12 827
Ontario, Ottawa	1 312	1 195	2 507	
Ontario, Sud-Est	1 294	1 080	2 374	
Ontario, Toronto	1 295	1 417	2 712	
Ontario, Reste	1 308	1 466	2 774	
Manitoba	1 269	1 299	2 568	2 568
Saskatchewan	1 166	1 426	2 592	2 592
Alberta	1 302	1 305	2 607	2 607
Colombie-Britannique	1 291	1 383	2 674	2 674
Whitehorse	332	391	723	723
Yellowknife	306	363	669	669
Iqaluit	251	203	454	454
Enquête sur la population de langue officielle en situation minoritaire	26 690	29 157	58 847	58 847

Source : Enquête sur la population de langue officielle en situation minoritaire de 2022.

4.0 Collecte de données

4.1 Période de collecte

Dans les 10 provinces, la collecte de données a eu lieu du 16 mai au 16 décembre 2022. Dans les trois capitales territoriales, la période de collecte a eu lieu du 22 août au 16 décembre 2022.

Comme il s'agit d'une enquête postcensitaire, la collecte a eu lieu à la suite du Recensement de la population de 2021, tenu le 11 mai 2021. Le Recensement a permis d'identifier la population de langue officielle en situation minoritaire qui est devenue la population cible pour l'EPLOSM.

4.2 Instrument et modes de collecte

Tel que mentionné auparavant, le contenu du questionnaire d'enquête a été élaboré à partir de consultations, entre autres, avec les membres d'un comité consultatif externe, ainsi qu'avec des partenaires fédéraux et communautaires. Le questionnaire électronique (QE) a ensuite été élaboré au moyen de vagues d'essais itératifs. Deux séries d'essais qualitatifs ont été menées dans les deux langues officielles.

Une fois que le questionnaire de l'EPLOSM a été finalisé, cinq modes de collecte ont été utilisés pour recueillir les données :

- Le questionnaire électronique à remplir par le répondant (QEr). Les répondants recevaient un code d'accès sécurisé leur permettant d'ouvrir une session et de répondre aux questions de l'enquête en ligne.
- Le questionnaire électronique administré par un interviewer (QEi), aussi connu comme l'interview téléphonique assistée par ordinateur (ITAO).
- Puisque diverses restrictions liées à la pandémie de COVID-19 étaient toujours en place au moment de la collecte, Statistique Canada a mis sur pieds la stratégie de collecte IPAO allégée plus (IAP). Les

intervieweurs se rendaient alors chez les répondants pour prendre un rendez-vous afin que soit rempli ultérieurement le questionnaire au moyen d'une ITAO.

- La stratégie de collecte '*Cogner, Parler et Appeler*' (CPA), semblable à l'IPAO allégée plus, mais avec laquelle l'ITAO a lieu immédiatement.
- Finalement, un mode de collecte pilote a été mis en place pendant environ un mois, à Iqaluit; les personnes sélectionnées ont été invitées à remplir leur questionnaire au moyen du QE-r ou de l'ITAO dans un immeuble municipal; des ordinateurs et du soutien en personne étaient offerts.

Les répondants ont pu choisir de remplir le questionnaire en français ou en anglais. En moyenne, il fallait environ 45 minutes pour le remplir.

4.2.1 Sécurité des questionnaires d'enquête en ligne

Le système électronique de collecte des données utilisé pour l'EPLOSM de 2022 faisait appel à un serveur Web sécurisé pour le QE en ligne, dans lequel on enregistrait les données d'enquête du QEr et du QEi.

Statistique Canada prend très au sérieux la protection des renseignements confidentiels transmis par les répondants, dont les renseignements fournis en ligne. Un processus sécurisé d'ouverture de session et un chiffrement robuste constituent des éléments essentiels pour empêcher quiconque de consulter ou de manipuler les données d'enquête des répondants provenant de questionnaires remplis et soumis en ligne.

Pour protéger la sécurité des renseignements personnels des répondants qui utilisent Internet, Statistique Canada a pris les mesures de sécurité suivantes :

- De robustes technologies de **cryptage bidirectionnel** ont été employées afin d'assurer la sécurité complète des données circulant entre l'ordinateur du répondant et le serveur Web de Statistique Canada.
- Les données de l'enquête ont été traitées et conservées dans un réseau interne à accès restreint auquel seules les personnes ayant prêté le serment de discrétion ont pu avoir accès. L'accès n'a été possible que selon le principe du « besoin de savoir ».
- Les données soumises aux serveurs Web de Statistique Canada ont été chiffrées avant d'être stockées, et elles le sont demeurées jusqu'à leur transfert au réseau interne hautement sécurisé.

De puissants pare-feu, un système de détection des intrusions et des procédures rigoureuses de **contrôle de l'accès** ont été utilisés pour limiter l'accès aux systèmes et aux bases de données.

4.3 Stratégie de collecte

Les personnes sélectionnées pour participer à l'EPLOSM ont reçu une lettre d'invitation par la poste qui décrivait l'enquête. Cette lettre les informait qu'elles avaient été sélectionnées pour participer et leur fournissait un code d'accès sécurisé pour ouvrir une session en ligne et répondre à l'enquête. Chaque lettre comprenait un lien à la page Web de l'EPLOSM en plus d'un numéro de téléphone sans frais pour les répondants préférant répondre au questionnaire à l'aide de l'ITAO ou ayant des questions touchant l'enquête (ainsi qu'un numéro ATS (appareil de télécommunications pour personnes sourdes) pour les personnes ayant une incapacité auditive). La lettre était aussi accompagnée d'un dépliant, en français et en anglais, qui décrivait l'enquête, l'importance de la participation et les sujets abordés. De plus, un dépliant en Inuktitut a été inclus pour les résidents d'Iqaluit.

Des rappels ont été envoyés par la poste et par message texte. Au cours de la période de collecte de sept mois, jusqu'à quatre lettres de rappel et six courriels de rappel ont été envoyés aux répondants qui n'avaient pas encore soumis leur questionnaire. Environ 80% des répondants ont rempli le QEr, les autres ayant rempli le QEi. On demandait aux intervieweurs de déployer tous les efforts raisonnables pour obtenir une interview complétée auprès du répondant sélectionné. Les personnes qui refusaient de participer ont reçu une lettre expliquant l'importance de l'enquête et les encourageant à y participer, et on communiquait à nouveau avec eux par téléphone. Dans la mesure du possible, les répondants d'Iqaluit ont été visités en personne par un employé de Statistique Canada qui a effectué le contact, expliqué l'enquête et pris un rendez-vous pour une interview le cas échéant. Partout au Canada, les interviews se sont déroulées dans la langue officielle choisie par le répondant, le français ou l'anglais.

4.4 Supervision et contrôle de la qualité

Tous les intervieweurs de Statistique Canada travaillaient sous la supervision d'intervieweurs principaux. Ces derniers devaient s'assurer que les intervieweurs connaissaient bien les concepts et les procédures de l'enquête qui leur a été confiée. Les intervieweurs principaux étaient aussi responsables de faire le suivi régulier du travail des intervieweurs.

Les intervieweurs ont reçu une formation sur le contenu de l'enquête, l'application ITAO et la façon de procéder avec les nouvelles stratégies de collecte IAP et CPA. En plus de la formation en ligne ou en classe, les intervieweurs ont mené une série d'interviews simulées afin de se familiariser avec l'enquête, ses concepts et ses définitions.

4.5 Interviews par personne interposée

La réponse par personne interposée (quand une personne remplit le questionnaire au nom d'une personne sélectionnée) n'était pas autorisée dans le cadre de l'EPLOSM. Par contre, pour l'échantillon des enfants, les parents étaient les répondants ciblés; ils remplissaient le questionnaire au nom de leur enfant sélectionné.

4.6 Stratégie de communication

Dans les mois précédant la collecte des données de l'EPLOSM, on a procédé à des préparatifs afin de fournir aux répondants tous les renseignements nécessaires à propos de l'enquête. Une page Web de l'EPLOSM a été affichée sur le site Web de Statistique Canada. La page Web de l'EPLOSM présentait des renseignements de base sur l'enquête et sa méthodologie, de même qu'un lien vers le questionnaire. De plus, on a créé une page Web spéciale de Renseignements pour les participants aux enquêtes (RPE) qui comprenait la marche à suivre pour participer, et une liste de « Questions et réponses » à propos de l'enquête.

Une campagne promotionnelle a été développée. Les partenaires communautaires ont été également invités à encourager la participation.

Des produits promotionnels incluant des dépliants et des feuillets d'information étaient disponibles en français et en anglais, et en Inuktitut à Iqaluit.

4.7 Situations particulières

4.7.1 Catastrophes naturelles

L'équipe de l'EPLOSM a surveillé de près la collecte afin de pouvoir réagir de façon proactive, au besoin, en cas de situation particulière. Par exemple, lors de l'ouragan Fiona, au cours duquel des états d'urgence locaux ont été déclarés dans plusieurs collectivités de la Nouvelle-Écosse, du Nouveau-Brunswick et de l'Île-du-Prince-Édouard en septembre 2022, la collecte a été interrompue dans les zones touchées afin de réduire le fardeau imposé aux collectivités qui tentaient de se remettre des dommages. Cela comprenait interrompre et réactiver les priorités téléphoniques pour le suivi ITAO à mesure que les collectivités se rétablissaient.

4.7.2 COVID-19

Les données de l'EPLOSM ont été recueillies deux ans après le début de la pandémie de COVID-19, pour une période d'environ sept mois. Bien que l'effet de la pandémie sur la collecte de données ne puisse être déterminé avec précision, cette situation doit être prise en considération. Des questions ont été ajoutées au questionnaire afin d'évaluer l'impact de la COVID-19 sur certaines des réponses.

4.8 Taux de réponse

Le taux de réponse a été calculé en divisant le nombre de réponses complètes par le nombre de personnes sélectionnées pour la collecte duquel on a soustrait les cas hors cible. Les cas hors cible comprennent, par exemple, les personnes qui, entre la tenue du Recensement de 2021 et la collecte des données de l'enquête, sont

décédées, qui ont émigré, qui ont été admises dans un établissement institutionnel ou qui ont déménagé dans une réserve des premières nations. De plus, les membres à temps plein des Forces canadiennes vivant sur une base militaire, les visiteurs au Canada (classés par erreur lors du recensement) étaient exclus de même que les répondants dont l'âge déclaré au moment de l'enquête était de moins de 15 ans.

Le taux de réponse est de 53,4%, soit 29 958 répondants. Le tableau suivant présente des taux de réponse (TR) et des comptes de répondants (#rép.) selon la province, la capitale territoriale ou la région, à la fois pour l'échantillon des adultes, celui des enfants et pour l'ensemble de l'enquête.

Tableau 3
Les taux de réponse (TR) et des comptes de répondants (#rép.) pour l'échantillon des adultes et des enfants et pour l'ensemble de l'enquête, selon la province, la capitale territoriale ou la région

	Adultes		Enfants		Région		Prov./Cap. (territoire)	
Région	**TR**	**#rép.**	**TR**	**#rép.**	**TR**	**#rép.**	**TR**	**#rép.**
Terre-Neuve-et-Labrador	41,7	345	50,6	484	46,5	829	46,5	829
Île-du-Prince-Édouard	40,4	402	50,8	457	45,3	859	45,3	859
Nouvelle-Écosse	47,6	556	55,7	797	52,1	1 353	52,1	1 353
Nouveau-Brunswick, Nord	44,0	556	49,8	481	46,5	1 037	51,4	3 447
Nouveau-Brunswick, Sud-Est	51,8	644	57,8	580	54,5	1 224		
Nouveau-Brunswick, Reste	49,8	570	56,4	616	53,0	1 186		
Québec, Est	53,0	567	45,7	455	49,5	1 022	58,5	11 219
Québec, Estrie et Sud	65,3	802	56,5	745	60,8	1 547		
Québec, Montréal	61,5	2 225	68,3	2 104	64,6	4 329		
Québec, Ouest	61,5	737	61,1	808	61,3	1 545		
Québec, Québec et environs	53,2	562	53,5	879	53,4	1 441		
Québec, Reste	53,7	631	47,4	704	50,2	1 335		
Ontario, Nord-Est	48,6	603	52,5	608	50,5	1 211	52,8	6 464
Ontario, Ottawa	53,0	653	59,4	705	56,1	1 358		
Ontario, Sud-Est	49,8	625	56,2	604	52,8	1 229		
Ontario, Toronto	45,4	503	60,0	844	53,5	1 347		
Ontario, Reste	42,9	489	57,2	830	50,9	1 319		
Manitoba	50,5	598	55,2	711	52,9	1 309	52,9	1 309
Saskatchewan	46,7	491	53,8	758	50,8	1 249	50,8	1 249
Alberta	42,7	501	54,7	704	49,0	1 205	49,0	1 205
Colombie-Britannique	43,3	487	54,7	743	49,6	1 230	49,6	1 230
Whitehorse	48,5	146	53,4	206	51,2	352	51,2	352
Yellowknife	42,8	122	51,0	184	47,4	306	47,4	306
Iqaluit	32,8	77	29,4	59	31,2	136	31,2	136
Total EPLOSM	**50,9**	**13 892**	**55,9**	**16 066**	**53,4**	**29 958**	**53,4**	**29 958**
Français (Hors-Québec)	46,6	8 368	54,8	10 371	50,8	18 739	...	...
Anglais (Québec)	59,1	5 524	57,8	5 695	58,5	11 219	...	...
Atlantique	46,3	3 073	53,8	3 415	50,0	6 488	...	...
Québec	59,1	5 524	57,8	5 695	58,5	11 219	...	...
Ontario	48,1	2 873	57,2	3 591	52,8	6 464	...	...
Prairies	46,6	1 590	54,5	2 173	50,9	3 763	...	...
Colombie-Britannique	43,3	487	54,7	743	49,6	1 230	...	...
Capitales des territoires	42,0	345	47,4	449	44,9	794	...	...
Provinces	51,1	13 547	56,2	15 617	52,0	29 164	...	...
Capitales des territoires	42,0	345	47,4	449	44,9	794	...	...

... n'ayant pas lieu de figurer
Source : Enquête sur la population de langue officielle en situation minoritaire de 2022.

5.0 Traitement des données

Le traitement transforme les réponses du questionnaire obtenues pendant la collecte pour qu'elles conviennent à la totalisation et à l'analyse des données. Il comprend toutes les activités de traitement des données, automatisées et manuelles, après la collecte et avant l'estimation.

5.1 Saisie des données

Les réponses aux questions de l'enquête ont été capturées dans le questionnaire éléctronique (QE), soit directement par le répondant (QEr) ou par l'intervieweur (QEi).

5.1.1 Questionnaire électronique à remplir par le répondant (QEr)

Les répondants sélectionnés en vue de participer à l'EPLOSM ont reçu un code d'accès sécurisé leur permettant d'ouvrir une session et de compléter l'enquête en ligne. Pour les questionnaires électroniques, les réponses aux questions de l'enquête ont été entrées directement par les répondants. Les réponses ont été sécurisées par le biais des protocoles de chiffrage selon les normes industrielles, des pare-feu et des couches de chiffrage.

5.1.2 Questionnaire électronique administré par un intervieweur (QEi)

Les intervieweurs saisissent directement les réponses aux questions de l'enquête au moment de l'interview à l'aide d'une version automatisée du questionnaire qui imite l'application du questionnaire électronique. De plus, les réponses ont été sécurisées par le biais des protocoles de chiffrage selon les normes industrielles, des pare-feu et des couches de chiffrage.

Pour le QEr et le QEi, une partie de la vérification se faisait au moment que le QE a été complété. Lorsque les renseignements introduits étaient hors limites (trop faibles ou trop élevés) des valeurs attendues, ou qu'ils entraient en contradiction avec des renseignements introduits auparavant, le répondant ou l'intervieweur voyaient s'afficher à l'écran de l'ordinateur des messages lui demandant de clarifier les renseignements. Cependant, pour certaines questions, le répondant ou l'intervieweur avaient la possibilité de contourner les vérifications et de sauter des questions si le répondant ne connaissait pas la réponse ou refusait de répondre. Pour cette raison, les données ont été soumises à d'autres processus de vérification après réception au bureau central. Une fois reçues, les données électroniques ont été converties en fichiers texte lisibles.

5.2 Étapes de traitement des enquêtes sociales

Le traitement des données comporte une série d'étapes pour convertir les réponses au questionnaire électronique de leur format brut à une base de données conviviale de grande qualité comprenant un ensemble exhaustif de variables pour l'analyse. De nombreuses opérations sont exécutées pour supprimer les erreurs accidentelles dans les fichiers, vérifier rigoureusement les données pour s'assurer la cohérence, coder les questions ouvertes, créer des variables utiles pour l'analyse des données et, enfin, systématiser et documenter les variables pour faciliter leur utilisation à des fins d'analyses.

Dans le cadre de l'EPLOSM de 2022, un ensemble d'outils de traitement des enquêtes sociales élaboré à Statistique Canada et appelé Environnement pour le traitement des enquêtes sociales (ETES) a été utilisé. L'ETES fait intervenir des programmes du logiciel SAS, des applications personnalisées et des processus manuels pour l'exécution des étapes systématiques suivantes :

- Réception des données brutes
- Nettoyage
- Recodage
- Codage
- Enchaînements

- Contrôle et imputation
- Couplage des données du recensement
- Variables dérivées
- Création d'un fichier de traitement final
- Validation des données
- Création de fichiers de diffusion
- Manuel de codage (dictionnaire de données)

5.3 Réception des données brutes et codage de données

Après la réception des données brutes de l'application du questionnaire électronique de l'EPLOSM, de nombreuses procédures de nettoyage préliminaire ont été mises en place au niveau des dossiers individuels. Cela comprenait la suppression de tous les identificateurs personnels des fichiers, comme les noms et les adresses, dans le cadre d'un ensemble rigoureux de mécanismes permanents visant à assurer la protection de la confidentialité des répondants. Les enregistrements en double ont été résolus à cette étape et les classifications de codage standard ont été appliquées. Par ailleurs, dans le cadre des procédures de nettoyage, tous les dossiers des répondants ont été révisés, afin de s'assurer que chaque répondant fasse partie du champ de l'enquête et avait un questionnaire suffisamment rempli. Il est important de souligner que les critères pour déterminer si un répondant faisait partie ou non du champ de l'enquête ont été appliqués avant tout contrôle ou imputation.

5.4 Enchaînements : cheminements de réponse, sauts valides et non-réponse à des questions

Un autre ensemble de procédures de traitement des données pour l'enquête comprenait la vérification des enchaînements des questions. Tous les cheminements de réponse et les enchaînements de questions intégrés au questionnaire ont été vérifiés, afin d'assurer que l'univers ou la population cible pour chaque question avait été saisi correctement lors du traitement. Une attention spéciale a été accordée aux distinctions entre les sauts valides et la non-réponse. Un saut valide veut dire qu'une question n'est pas pertinente en raison d'une réponse particulière à une question précédente, et des cheminements ont été construits pour qu'on ne pose pas la question au répondant. Le répondant n'a donc jamais vu la question, alors celle-ci est codée comme saut valide. Si le répondant a vu la question et n'a pas répondu, la question est codée comme « Non déclaré ».

Des codes spéciaux ont été désignés pour chacun de ces types de réponses, afin de faciliter la reconnaissance et l'analyse des données par l'utilisateur. Par exemple, le dernier chiffre des codes de « saut valide » est « 6 », et les chiffres précédents sont des « 9 » (p. ex. le code serait « 996 » pour une variable à trois chiffres). Toutes les réponses « Ne sais pas » se terminent par un « 7 », et sont précédées de « 9 » (p. ex. « 997 »). Toutes les réponses « Non déclaré » se terminent par un « 9 », et sont précédées de « 9 » (p. ex. « 999 »). Ces codes de réserve sont différents pour chaque variable, selon le nombre de catégories que comprend la variable et la longueur de la variable.

5.5. Codage des questions ouvertes

5.5.1 Questions « Autre — Précisez »

Le traitement des données comprend aussi le codage des réponses écrites aux questions, aussi appelées « Autre — Précisez ». Lorsque la réponse d'un répondant ne pouvait être facilement attribuée à une catégorie existante, de nombreuses questions permettaient au répondant ou à l'intervieweur de saisir une réponse en toutes lettres dans la catégorie « Autre — Précisez ».

Toutes les questions comportant la catégorie « Autre — Précisez » ont fait l'objet d'un examen minutieux au cours du traitement. À la suite d'un examen qualitatif des types de réponses écrites fournies, des lignes directrices de codage ont été élaborées pour chaque question. À partir de ces lignes directrices de codage, bon nombre des réponses écrites fournies ont été codées à nouveau dans l'une des catégories de réponses préexistantes.

Parfois, une ou plusieurs nouvelles catégories peuvent être créées lorsqu'il y a un nombre suffisant de réponses pour les justifier, mais ce n'était jamais le cas pour l'EPLOSM. Les réponses qui étaient uniques et différentes des catégories existantes ont été conservées dans la catégorie « Autres ».

5.5.2 Questions ouvertes et classifications types

Les réponses à certaines questions de l'enquête ont été consignées dans un format complètement ouvert. Celles-ci comprenaient des questions liées à la profession et à l'industrie de travail du répondant, le cas échéant. Ces réponses ont été codées à l'aide d'une combinaison de procédures de codage automatisées et interactives. Des systèmes de classification normalisés ont été utilisés pour coder ses réponses.

5.6 Vérification

Des vérifications peuvent être faites à plusieurs étapes pendant le processus de l'enquête et elles passent des simples vérifications préliminaires dans l'application de collecte électronique aux vérifications automatisées plus complexes exécutées pendant le traitement après la saisie des données. Les règles de la vérification sont généralement formulées selon ce qui peut être logique ou valide, compte tenu :

- des connaissances de l'expert en la matière,
- d'autres enquêtes ou données connexes,
- de la structure du questionnaire et de ses questions,
- d'une théorie statistique.

Il existe trois grandes catégories de vérification : vérification de validité, de cohérence et de distribution.

Les vérifications de validité ciblent la syntaxe des réponses et peuvent notamment permettre de déterminer que les données s'inscrivent dans l'étendue permise des valeurs. Une vérification de l'étendue peut être faite, par exemple, pour l'âge déclaré d'un répondant, afin de vérifier s'il se situe entre 0 et 121 ans.

Les vérifications de cohérence déterminent si les liens entre les questions sont respectés. Les vérifications de cohérence peuvent utiliser des liens logiques, juridiques, comptables ou structurels entre les questions ou entre les volets d'une question. Le lien entre la date de naissance et l'état matrimonial est un exemple auquel la vérification de cohérence peut être appliquée : « l'état matrimonial d'une personne de moins de 15 ans peut seulement être jamais marié ».

Les vérifications de distribution sont faites en observant les données entre les questionnaires. Elles tentent de déterminer les enregistrements qui sont des valeurs aberrantes du point de vue de la distribution des données. Les vérifications de distribution sont parfois considérées comme des vérifications statistiques ou des détections de valeurs aberrantes.

5.7 Imputation

Comme décrit dans la section 6, la non-réponse totale a été comblée par la repondération des répondants. Aucune imputation n'a été faite pour la non-réponse partielle, pour convertir des valeurs manquantes aux variables d'enquête à une de ces réponses admissibles.

5.8 Couplage aux données du recensement (addition des variables du recensement)

Étant donné que l'échantillon de l'EPLOSM a été tiré de la base de données du recensement, les réponses à certaines questions et variables pertinentes du recensement ont été ajoutées à la base de données de l'EPLOSM par l'appariement des données. Ces variables sont identifiées, dans le dictionnaire des données, comme variables du Recensement de la population de 2021, sous la ligne de source d'une variable donnée. En outre, environ 250 variables du Recensement de 2021 ont été couplées au fichier analytique final de l'EPLOSM de 2022. Pour

toutes les variables couplées du recensement, le nom de la variable du recensement a été préservé dans la mesure du possible dans la base de données de l'EPLOSM. Toutefois, certaines variables du recensement ont été renommées, car les noms des variables de l'EPLOSM se limitent à huit caractères, alors que les noms des variables du recensement dépassent parfois cette limite. Dans ces cas, une note indiquant le nom original de la variable du recensement à partir duquel la variable a été abrégée a été ajoutée au dictionnaire de données.

5.9 Création de variables dérivées

Pour faciliter l'analyse des données, un certain nombre de données élémentaires incluses dans le fichier de microdonnées ont été calculées en combinant des éléments du questionnaire. Ainsi, les catégories de réponses d'une, de deux ou de plusieurs variables peuvent être combinées pour créer une nouvelle variable. Les variables dérivées peuvent être elles-mêmes utilisées comme base pour d'autres variables dérivées.

Par exemple, la variable 'Première langue officielle parlée du répondant (PLOPRESP) a été dérivée des réponses aux questions sur la connaissance des langues officielles, la langue maternelle et les langues parlées à la maison.

Le dictionnaire de données précise les variables qui ont été dérivées, ainsi que les variables de base sur lesquelles ces variables dérivées sont basées.

Afin de faciliter une analyse plus approfondie du riche ensemble de données de l'EPLOSM, environ 200 variables dérivées ont été créées en regroupant des questions du questionnaire. Des variables dérivées (VD) ont été créées pour tous les principaux domaines de contenu.

Quelques VD simples consistaient à fusionner des catégories pour créer des catégories plus générales. Dans d'autres cas, deux variables ou plus ont été regroupées pour créer une variable nouvelle ou plus complexe, qui serait utile pour les analystes des données. Certaines des VD étaient fondées sur des variables couplées du recensement, alors que les données d'un répondant qui refuse le couplage avec le recensement sont supprimées des variables du recensement et des variables fondées sur le recensement.

Pour la plupart des VD, il y a une catégorie résiduelle étiquetée « Non déclaré» pour les cas où les réponses aux VD des questions sources ne satisfont pas aux conditions pour placer un répondant dans l'une ou l'autre des catégories valides de la VD. Dans bien des cas, mais pas tous, un répondant est inclus dans la catégorie « Non déclaré» si l'une ou l'autre des parties de l'équation n'avait pas reçu de réponse (c'est-à-dire si l'une ou l'autre des questions utilisées pour la VD avait été codée « Ne sais pas » ou « Non déclaré»).

Le dictionnaire de données de l'EPLOSM de 2022 indique de façon détaillée les variables qui ont été dérivées et précise à partir de quelles variables sources elles ont été dérivées. Une liste complète des variables couplées du recensement et de leurs notes figure dans le dictionnaire de données de l'EPLOSM de 2022 qui accompagne le fichier analytique de l'enquête.

5.10 Contrôle de la divulgation

La Loi interdit à Statistique Canada de divulguer toute information que l'organisme recueille qui pourrait permettre d'identifier une personne, une entreprise ou une organisation, sauf si le répondant a donné son consentement ou si la Loi sur la statistique le lui permet. 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.

Des instructions spécifiques sur la façon de publier des informations statistiques à partir des données de l'EPLOSM afin de satisfaire à ces exigences sont fournies dans la section 8.

5.11 Création de fichiers de données finaux et d'un dictionnaire de données (manuel de codage)

Deux fichiers de données finaux ont été créés, un pour les adultes et un pour les enfants. Les fichiers analytiques ont été traités au complet et préparés pour la diffusion.

Afin de transformer le fichier de traitement final épuré en fichier analytique pour les chercheurs, un certain nombre d'étapes ont été suivies. Tout d'abord, une série de mesures ont été appliquées pour améliorer la protection de la confidentialité des répondants. Ensuite, des poids-personne ont été ajoutés au fichier. Enfin, toutes les variables temporaires ou variables utilisées exclusivement pour le traitement ont été supprimées du fichier de traitement final.

Le fichier analytique de l'EPLOSM de 2022 comprend le cliché d'enregistrement, la syntaxe SAS, SPSS et Stata servant à télécharger le fichier, ainsi que des métadonnées sous forme d'un dictionnaire de données qui décrit chaque variable et fournit des fréquences pondérées et non pondérées.

Les deux fichiers analytiques ont été créés afin d'être utilisés dans les centres de données de recherche (CDR) et avec le service d'accès à distance en temps réel (ADTR). Les fichiers analytiques sont disponibles dans les CDR partout au Canada, mais sont accessibles uniquement pour les chercheurs qui répondent à certaines exigences. Les chercheurs associés à un établissement institutionnel, un ministère ou un organisme sans but lucratif peuvent également utiliser l'outil d'ADTR pour avoir accès au fichier. Le fichier analytique est aussi utilisé à Statistique Canada pour produire des tableaux de données en réponse aux demandes des clients.

Centres de données de recherche

Le Programme des CDR permet aux chercheurs et chercheuses d'utiliser les données d'enquête des fichiers maîtres dans un environnement sécurisé situé dans plusieurs universités au Canada. Les chercheurs et chercheuses doivent soumettre des propositions de recherche, qui, une fois acceptées, leur donneront accès à un CDR.

Voir la page web suivante pour plus de renseignements : Centres de données de recherche

Accès à distance en temps réel

Le système d'ADTR est une installation d'accès à distance en ligne qui permet aux utilisateurs d'exécuter des programmes SAS, en temps réel, portant sur des ensembles de microdonnées situés dans un emplacement central et sécurisé. Les chercheurs et chercheuses soumettent des programmes SAS pour extraire les résultats sous la forme de tableaux de fréquences. La confidentialité des microdonnées est assurée dans le système d'ADTR par la suppression des variables du fichier maitre ne comportant pas suffisamment d'enregistrement afin d'assurer le respect de la confidentialité. Les chercheurs et chercheuses doivent remplir un formulaire de demande

Voir la page web suivante pour plus de renseignements : Système d'accès à distance en temps réel

Totalisations personnalisées

Une autre façon d'accéder aux fichiers maîtres consiste à offrir à tous les utilisateurs la possibilité de demander aux Services à la clientèle de l'EPLOSM de préparer des totalisations personnalisées. Grâce à ce produit à recouvrement des coûts, les utilisateurs qui ne maîtrisent pas les logiciels de totalisation peuvent obtenir des résultats personnalisés. Les résultats sont examinés 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 : statcan.languagestats-statlinguistique.statcan@statcan.gc.ca.

6.0 Erreurs d'enquête

L'erreur d'enquête est la différence numérique entre la valeur inconnue du paramètre de population d'intérêt et son estimation obtenue des données. Diverses sources expliquent les erreurs d'enquête. Elles peuvent être classées en deux principales catégories : erreur d'échantillonnage et erreurs non dues à l'échantillonnage.

6.1 Erreurs non dues à l'échantillonnage

Les erreurs non dues à l'échantillonnage sont les erreurs résultant de toute activité d'enquête à l'exception de l'échantillonnage. Ces erreurs se retrouvent dans l'enquête-échantillon et le recensement (contrairement à l'erreur d'échantillonnage qui est présente seulement dans l'enquête-échantillon). Les principales sources d'erreurs non dues à l'échantillonnage sont : non-réponse, couverture, mesure et traitement. Par exemple, une question dont la signification n'est pas exactement la même en français et en anglais est source d'erreur non due à l'échantillonnage.

6.1.1 Non-réponse

La non-réponse résulte de l'incapacité à recueillir l'information nécessaire pour toutes les unités sélectionnées dans l'échantillon. Non seulement la non-réponse implique moins de données à analyser pour les utilisateurs, mais les non-répondants peuvent également avoir des caractéristiques différentes de celles des répondants, ce qui peut entraîner des estimations d'enquête biaisées.

La non-réponse est généralement classée soit comme non-réponse totale, soit comme non-réponse partielle. La non-réponse est totale lorsque toutes[7] les informations nécessaires à l'enquête à propos d'une personne échantillonnée sont manquantes ou inutilisables; la non-réponse est partielle lorsque seulement certaines de ces informations sont manquantes ou inutilisables.

Les pratiques exemplaires d'enquête ont été employées avant, pendant et après la collecte de données afin de réduire la non-réponse au sein de l'EPLOSM. Par exemple, avant le début de la collecte, des campagnes médiatiques ont été réalisées et des brochures attrayantes ont été distribuées pour informer les communautés ciblées par l'EPLOSM. Et pendant le déroulement de la collecte, un suivi constant de la non-réponse a été effectué afin de rediriger les efforts et les ressources là où ils étaient le plus nécessaire.

Pour un domaine donné, le taux de réponse de l'EPLOSM a été obtenu en divisant le nombre de personnes répondantes par le nombre de personnes sélectionnées qui appartiennent au champ de l'enquête; le taux de réponse global est 53,4 % et il est comparable à celui des autres enquêtes postcensitaires qui étaient actives au même moment. Le tableau de la section 4.8 décrit le taux de réponse obtenu pour divers domaines d'intérêt; la prochaine section traite des personnes hors du champ de l'enquête.

Pour l'EPLOSM, la non-réponse partielle n'a pas été comblée en recourant à l'imputation : les valeurs manquantes aux variables d'enquête demeurent dans le fichier analytique sous la forme « non déclaré ». Dans les enquêtes sociales du type de l'EPLOSM, on procède par repondération afin d'atténuer les effets de la non-réponse totale sur les estimations – voir la section 7.1 pour de plus amples détails.

6.1.2 Erreurs de couverture

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

Certaines des personnes adultes faisant initialement partie du champ de l'enquête selon les données du recensement se sont révélées ne pas en faire partie sur la base de leurs réponses à l'EPLOSM, et ce pour diverses raisons dont : 1) elles ne faisaient plus partie de la population de langue officielle en situation minoritaire lors du 16

7. Une personne peut avoir répondu à certaines questions, telles celles du module d'identification seulement, et être considérée non-répondante. Une personne doit avoir fourni des réponses valides à un ensemble prédéterminé de questions clés de l'enquête (appelé « contenu minimum ») afin d'être considérée répondante.

mai 2022, un an après le recensement, 2) elles ont déménagé à l'extérieur du Canada, 3) elles vivent maintenant en établissement, 4) elles sont décédées. Les mêmes raisons s'appliquent aux enfants, mais il n'était pas possible d'un point de vue opérationnel de les évaluer puisque l'admissibilité d'un enfant à l'enquête est établie en fonction du profil linguistique de ses parents et non juste de l'adulte qui répond en son nom.

Le taux de personnes hors du champ de l'enquête pour un domaine donné, exprimé en points de pourcentage, a été obtenu en divisant le nombre de personnes hors champ du domaine par le nombre de personnes échantillonnées du domaine auprès desquelles un contact a été établi lors de la collecte.

En ce qui concerne les adultes, le changement de leur statut à l'égard de la population de langue officielle en situation minoritaire explique 80 % (une statistique non présentée dans le tableau) de tous les cas de personnes hors du champ de l'enquête. Dans la plupart de ces cas, il s'agit de personnes qui selon les données du recensement parlaient les deux langues officielles, mais qui ont rapporté à l'EPLOSM ne pas parler la langue officielle en situation minoritaire. Cela peut s'expliquer en partie par la déclaration par procuration qui amène une même personne à répondre aux questions de langue du recensement au nom de toutes les autres personnes de son ménage sans connaître exactement leur profil linguistique.

Tableau 6.1
Taux de personnes hors du champ de l'enquête : global et excluant les adultes dont le statut de langue officielle en situation minoritaire a changé

Région	Adultes	Enfants	Région[1]	Prov./Cap. de terr[1]
	pourcentage			
Terre-Neuve-et-Labrador	16,4	1,0	8,4 (1,2)	8,4 (1,2)
Île-du-Prince-Édouard	9,3	2,1	5,9 (1,4)	5,9 (1,4)
Nouvelle-Écosse	9,5	1,5	5,1 (1,3)	5,1 (1,3)
Nouveau-Brunswick Nord	3,6	0,8	2,3 (1,2)	3,7 (1,2)
Nouveau-Brunswick Sud-Est	4,9	0,4	2,9 (1,1)	
Nouveau-Brunswick Reste	9,7	1,5	5,7 (1,4)	
Québec, Est	4,3	3,6	4,0 (2,5)	5,7 (2,0)
Québec, Estrie et Sud	4,7	2,2	3,4 (1,6)	
Québec, Montréal	9,2	1,3	5,7 (1,8)	
Québec, Ouest	8,5	2,7	5,5 (2,0)	
Québec, Québec et environs	18,7	3,2	9,8 (2,7)	
Québec, Reste	8,2	2,9	5,4 (1,9)	
Ontario Nord-Est	5,7	0,7	3,2 (1,1)	5,7 (1,1)
Ontario, Ottawa	7,9	0,8	4,4 (0,8)	
Ontario Sud-Est	4,0	0,7	2,5 (0,9)	
Ontario, Toronto	19,1	0,9	9,1 (1,0)	
Ontario, Reste	17,2	1,3	8,5 (1,5)	
Manitoba	8,9	1,0	4,7 (1,0)	4,7 (1,0)
Saskatchewan	13,0	1,5	6,4 (1,5)	6,4 (1,5)
Alberta	14,0	1,6	7,4 (1,6)	7,4 (1,6)
Colombie-Britannique	17,5	2,2	9,2 (1,9)	9,2 (1,9)
Whitehorse	10,4	1,4	5,5 (1,9)	5,5 (1,9)
Yellowknife	7,8	0,6	3,8 (1,0)	3,8 (1,0)
Iqaluit	7,5	1,1	4,6 (0,8)	4,6 (0,8)
EPLOSM	**10,0**	**1,6**	**5,7 (1,5)**	**5,7 (1,5)**

1. Deux taux sont rapportés au niveau des régions linguistiques et des provinces/capitales des territoires : le taux global et, entre parenthèses, le taux qui exclut les changements de langue officielle minoritaire rapportés par certains adultes – c'est le premier facteur ci-dessus. De ces taux on peut voir que les taux pour les adultes hors du champ de l'enquête se comparent à ceux des enfants une fois ce facteur exclut de l'analyse.

Source : Enquête sur la population de langue officielle en situation minoritaire de 2022.

Tel que mentionné dans la Section 3.2, les enfants sont classés dans l'un des nombreux groupes en fonction du profil linguistique et du profil scolaire de leurs parents. Pour des raisons opérationnelles, l'échantillonnage a été mené à partir d'une classification préliminaire des enfants à partir des données du recensement. Par la suite, lorsque la classification a été complétée, des enfants ont dû être reclassifiés et leur poids d'enquête ajusté.

6.1.3 Erreurs de mesure

Les erreurs de mesure (ou parfois appelées erreurs de réponse) sont la différence entre la réponse inscrite à une question et la « vraie » valeur. Le répondant, l'intervieweur, le questionnaire, la méthode de collecte ou le système de conservation des dossiers du répondant peuvent susciter ce genre d'erreur. De telles erreurs peuvent être aléatoires ou peuvent entraîner un biais systématique si elles ne sont pas aléatoires.

Les questions sélectionnées pour une possible inclusion dans le questionnaire de l'EPLOSM de 2022 ont par la suite fait l'objet de plusieurs rondes d'essais qualitatifs au moyen d'interviews virtuelles individuelles avec les répondants dans différentes communautés partout au Canada. Des essais qualitatifs du questionnaire d'enquête ont été menés par le Centre de ressources en conception de questionnaires (CRCQ) de Statistique Canada. À partir des résultats de ces essais, des ajustements ont été apportés au libellé des questions et aux enchaînements en vue de réduire l'erreur de mesure.

Le questionnaire EPLOSM a aussi utilisé le contenu harmonisé – des questions fréquemment posées, formatées de manière standard pour qu'elles puissent être utilisées dans plusieurs enquêtes.

De nombreux autres moyens ont aussi été pris pour réduire spécifiquement l'erreur de mesure, y compris le recours à des intervieweurs qualifiés, la formation exhaustive des intervieweurs en ce qui a trait aux procédures et au contenu de l'enquête, ainsi que l'observation et la supervision des intervieweurs, afin de déceler les problèmes de conception du questionnaire ou d'une mauvaise compréhension des instructions. De surcroît, les données entrantes ont été évaluées en temps réel afin de détecter les problèmes de conception du questionnaire ou de mauvaise compréhension des instructions.

Il est très onéreux de mesurer avec précision le niveau de l'erreur de réponse et très peu d'enquêtes mènent ce genre d'étude. Cependant, la rétroaction des intervieweurs et les rapports d'observation fournissent souvent des pistes pouvant mener aux questions plus problématiques (mauvais libellé, instructions aux intervieweurs insuffisantes, mauvaise traduction, jargon technique, pas de texte d'aide disponible, etc.).

6.1.4 Erreurs de traitement

L'erreur de traitement est associée aux activités suivant la réception des réponses d'enquête. Elles incluent toutes les activités de manipulation de données après la collecte et avant l'estimation. Elles peuvent être aléatoires comme toutes les autres erreurs et accroître ainsi la variance des estimations de l'enquête, ou elles peuvent être systématiques et ajouter un biais. Il est souvent difficile d'obtenir des mesures directes des erreurs de traitement ainsi que leur impact sur la qualité des données puisqu'elles sont souvent confondues avec d'autres types d'erreurs (non-réponse, mesure et couverture).

Des erreurs de traitement peuvent se produire à diverses étapes du processus d'enquête, notamment lors de la programmation du questionnaire électronique, lors de la saisie des réponses par l'intervieweur ou par le répondant lui-même, lors du codage et lors de la vérification des données. Des procédures de contrôle de la qualité ont été appliquées à chaque étape du traitement des données de l'EPLOSM afin de réduire ce type d'erreur. La collecte des données a été réalisée au moyen d'un questionnaire électronique, soit administré par un intervieweur, soit rempli en ligne par le répondant. Un certain nombre de vérifications ont été intégrées au système afin d'avertir le répondant ou l'intervieweur en cas d'incohérence ou de valeurs inhabituelles, permettant ainsi de corriger ces incohérences ou erreurs immédiatement (voir la section 5.6).

À l'étape du traitement des données, une série précise de procédures et de règles de vérification a été utilisée afin de repérer et de corriger les incohérences entre les réponses fournies. Pour chaque étape du nettoyage des données, un ensemble de procédures systématiques complètes a été mis au point afin d'évaluer la qualité de chaque variable du fichier et d'apporter des corrections si nécessaire. Un aperçu des fichiers de sortie a été établi à chaque étape, et une vérification a été effectuée en comparant les fichiers à l'étape en cours et à l'étape précédente. La programmation de toutes les règles de vérification a fait l'objet d'essais avant d'être appliquée aux données. À titre d'exemples de vérification du traitement des données, mentionnons :

- l'examen de l'enchaînement des questions, y compris les séquences très complexes, afin de faire en sorte que les valeurs de sauts de questions ont été attribuées avec exactitude et qu'elles se distinguent des différents types de valeurs manquantes;
- l'examen qualitatif assidu des questions ouvertes et des réponses « Autre – Précisez » afin d'assurer un codage approprié et robuste;
- les opérations de codage par rapport aux classifications types;
- l'examen des taux de non-réponse au niveau de la question;
- l'examen des variables dérivées par rapport aux variables de leur composante, afin d'assurer la programmation adéquate de la logique de calcul, y compris les calculs très complexes;
- lorsque cela était possible, une comparaison à l'enquête de 2006, et aux autres sources de données, a été fait.

Voir la section sur le traitement des données du présent guide pour obtenir plus de détails (section 5).

6.2 Erreur d'échantillonnage

L'erreur d'échantillonnage est définie comme l'erreur résultant de l'estimation d'une caractéristique de la population à partir des données collectées auprès d'un échantillon seulement des individus d'intérêt. Bien que l'erreur d'échantillonnage puisse se manifester sous forme de biais et de variance, des pratiques exemplaires et des méthodes éprouvées ont été employées afin de minimiser le biais de sorte que seule la variance soit à évaluer. Pour l'EPLOSM, une méthode bootstrap adaptée aux caractéristiques propres de son plan d'enquête est employée pour estimer la variance des estimations.

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

1. La variabilité de la caractéristique d'intérêt dans la population : plus la caractéristique dans la population est variable, plus la variance est grande.
2. La taille de la population : en général, la taille de la population a des répercussions sur la variance seulement pour les populations de petite taille ou de taille moyenne.
3. La taille de l'échantillon : toutes autres choses étant égales, la variance décroît avec un accroissement de la taille d'échantillon. (Cela représente donc une façon pratique de réduire l'erreur totale d'enquête.)
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 mener à une variance moindre que d'autres. Dans certaines situations, des avantages opérationnels peuvent justifier le recours à un plan de moindre efficacité.

Bien que la variance saisisse la totalité des erreurs dues à l'échantillonnage au sein d'estimations sans biais, elle n'est pas utilisée directement pour rapporter la précision d'une estimation car l'erreur-type, le coefficient de variation (CV) et l'intervalle de confiance (IC) – tous des sous-produits de la variance – lui sont plus utiles et informatifs.

L'erreur-type[8] d'un estimateur est la racine carrée de sa variance. Cette mesure est plus facile à interpréter parce qu'elle utilise la même échelle que l'estimation elle-même. Par exemple, l'erreur-type d'une estimation de revenus s'exprimera elle aussi en dollars, alors que la variance aura pour unités des dollars au carré (ou dollars quadratiques). Cependant, l'erreur-type ne se prête pas toujours à des comparaisons de précision entre estimations. Par exemple, lorsque l'effectif total N de la population est connu, diviser l'estimation d'un total d'intérêt par 1 (ce qui laisse l'estimation inchangée) ou par N (pour en tirer une estimation de la moyenne d'intérêt correspondante) n'a aucun impact sur la précision de ces estimations puisque ce sont là des constantes. Or, l'erreur-type de la moyenne estimée sera N fois plus petite que celle de l'estimation du total associé, laissant croire que la première est plus précise que la seconde.

8. Ce terme est souvent employé de façon interchangeable avec « écart-type », un terme apparenté mais non équivalent. L'écart-type est la racine carrée d'une variance associée. L'erreur-type d'un estimateur est l'écart-type de sa variance.

Le CV d'une estimation est une mesure relative de l'erreur. Il est calculé en divisant l'erreur-type par l'estimation elle-même. Sans dimension, il est habituellement exprimé en pourcentage (10 % au lieu de 0,1). Le CV est utile pour mesurer et comparer l'erreur de variables quantitatives avec de grandes valeurs positives. En retournant à l'exemple précédent, on peut voir que les CV des estimations du total et de la moyenne correspondante seraient les mêmes. Cependant, le CV n'est pas recommandé d'emploi pour des estimations telles que les proportions, les estimations de changements ou de différences, et les variables qui peuvent avoir des valeurs négatives.

Il est considéré comme une pratique exemplaire à Statistique Canada de faire état de l'erreur d'une estimation par l'entremise 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, selon la formulation bien connue employée par les sondages d'opinion), l'intervalle de confiance couvrirait la véritable valeur dans la population.

C'est là l'interprétation probabiliste d'un IC. Cependant, le lecteur pourra trouver l'argument heuristique suivant plus révélateur. En raison de l'erreur d'enquête, il n'est pas possible de dire avec certitude si l'estimation obtenue des données est proche ou non de la vraie valeur. En d'autres mots, bien que cette estimation représente une valeur plausible selon les données pour le paramètre de population d'intérêt, il y a toute une gamme d'autres valeurs avoisinantes à l'estimation qui soient également plausibles. Cette gamme de valeurs plausibles constitue l'IC de niveau 95 %. Plus court est l'IC, plus petite est la gamme de valeurs plausibles (incluant l'estimation) pour le paramètre de population d'intérêt, et donc plus vraisemblable est le fait que l'estimation soit proche de la vraie valeur.

L'annexe D fournit des renseignements généraux à propos des intervalles de confiance dans le contexte des enquêtes postcensitaires, alors que la section 8.6 décrit comment obtenir un intervalle de confiance valide de niveau 95 % à partir des données de l'EPLOSM.

7.0 Poids d'enquête et poids Bootstrap

Le principe qui sous-tend l'estimation pour un échantillon probabiliste veut que chaque unité incluse dans l'échantillon représente, en plus d'elle-même, plusieurs autres unités non sélectionnées dans l'échantillon. Par exemple, dans un échantillon aléatoire simple de taille 100 sélectionné d'une population de taille 5 000, chaque unité incluse dans l'échantillon représente 50 unités de la population (incluant elle-même). Le nombre d'unités représentées par une unité sélectionnée dans l'échantillon constitue le poids d'enquête de l'unité sélectionnée.

La présente section renferme des détails au sujet de la méthode utilisée pour calculer les poids d'enquête pour l'EPLOSM. Le poids d'enquête s'emploie pour réduire le biais que les activités d'enquête (p. ex., l'échantillonnage et la non-réponse) sont susceptibles d'introduire dans les estimations. Les 1 000 poids bootstrap s'utilisent afin d'estimer la variance d'une estimation pondérée qu'on rapportera sous la forme d'un IC. La section 8.6 contient des instructions détaillées quant à l'usage des poids bootstrap.

7.1 La variable de poids d'enquête

Puisque l'EPLOSM est une enquête postcensitaire, son plan d'échantillonnage comporte plusieurs phases dont la sélection d'un échantillon aléatoire simple stratifié à partir de la base de sondage constituée à partir des réponses obtenues lors du recensement. Par conséquent, une variable de poids d'échantillonnage a été dérivée à partir de la variable de poids du recensement et des fractions de sondage de l'EPLOSM.

Pour diverses raisons de nature opérationnelle, certains individus de l'échantillon ont été retirés des activités de collecte de données, incluant les cas où :

1. plus de trois personnes d'un même ménage ont été sélectionnées par l'EPLOSM
2. plus de trois personnes d'un même ménage ont été sélectionnées lorsque les deux autres enquêtes postcensitaires (l'Enquête canadienne sur l'incapacité et l'Enquête auprès des peuples autochtones) ont été prises en compte
3. les échantillons de l'EPLOSM et de l'Enquête auprès des peuples autochtones se recoupent

4. l'information disponible pour la prise de contact est insuffisante pour soutenir la collecte de données
5. de multiples enregistrements sont associés à une même personne en analysant les noms, dates de naissance et adresses qu'ils comportent.

Dans les deux premiers cas, les individus affectés échantillonnés ont été retirés aléatoirement, et leur poids d'échantillonnage redistribué parmi ceux restants. Dans les autres cas, les individus concernés (ou les enregistrements excédentaires) ont été retirés et leur poids d'échantillonnage redistribué parmi ceux restants.

La repondération effectuée au sein de l'EPLOSM afin de corriger la non-réponse totale suppose que la non-réponse, qu'on se représente comme un mécanisme aléatoire, s'explique entièrement à partir des informations disponibles à l'équipe d'enquête à propos des personnes répondantes et non-répondantes. En d'autres mots, on suppose que la non-réponse ne dépend pas de facteurs non mesurés. C'est pourquoi l'échantillonnage de l'EPLOSM a été mené autant que possible à partir des données du questionnaire détaillé du recensement plutôt que celles du questionnaire abrégé, afin de disposer d'informations sociodémographiques détaillées communes à propos des personnes répondantes et non-répondantes à l'EPLOSM.

Du point de vue des opérations, trois cas ont été traités séparément et successivement : 1) les personnes non contactées, 2) la non-réponse totale à la suite d'un contact, et 3) non-réponse partielle à la suite d'un contact. Le premier cas est composé de personnes qui n'ont pas pu être contactées en dépit des informations dont dispose le recensement à leur sujet et des meilleurs efforts déployés par l'équipe d'enquête. Leur poids a été partagé parmi les personnes auprès desquelles un contact a pu être établi. La non-réponse totale survient lorsqu'une personne échantillonnée a pu être contactée, mais n'a pas fourni les informations clés demandées par l'enquête. Dans ce cas, le poids des personnes non-répondantes a été partagé parmi les personnes ayant répondu. Finalement, les répondants partiels correspondent aux personnes n'ayant pas fourni tous les renseignements nécessaires afin qu'elles soient considérées répondantes. Ces cas ont été convertis en non-réponse totale.

Les ajustements pour tenir compte de la non-réponse ont été effectués en trois étapes. À la première étape, un modèle de régression logistique a été utilisé pour prédire la probabilité de répondre (probabilité d'obtenir une réponse) pour chaque unité (à la fois pour les unités répondantes et non-répondantes), à partir d'une série de variables explicatives. Les variables explicatives sont des caractéristiques « personnes » ou « ménages » provenant du Recensement de 2021 (p. ex. le nombre de personnes dans le ménage de la personne sélectionnée) pour la personne sélectionnée ou la personne la mieux renseignée dans le cas des enfants.

Dans le cadre de la deuxième étape, les répondants et non-répondants ayant des probabilités prédites de réponse semblables ont été répartis en classes d'ajustement en se servant de l'analyse de classification. Le taux de réponse a été calculé pour chacune des classes en fonction du nombre de répondants et de non-répondants dans la classe. Le taux de réponse calculé a ensuite été pondéré à partir des poids obtenus à l'étape d'ajustement précédente.

Dans le cadre de la troisième étape, les poids des unités répondantes à l'intérieur de chaque classe ont été ajustés à l'aide de l'inverse du taux de réponse pondéré de cette catégorie. Le poids des personnes qui n'ont pas répondu a été mis à zéro.

À l'étape finale de création de la variable de poids d'enquête, une post-stratification a été effectuée selon la région employée lors de la stratification, le groupe d'âge et le groupe linguistique établis à partir des informations mises à jour afin d'aligner les estimations correspondantes avec celles issues du recensement. Le poids d'enquête a été corrigé lorsqu'il contenait des valeurs jugées trop extrêmes.

La variable de poids d'enquête de l'EPLOSM s'appelle WEIGHT.

Le plan d'échantillonnage de l'EPLOSM est complexe, ce qui se traduit par une variable de poids d'enquête qui prend des valeurs différentes d'une personne à une autre au sein du fichier analytique. Afin de produire des estimations et des tableaux statistiques valides à partir des données de l'EPLOSM, les analystes doivent utiliser la variable de poids d'enquête. Ne pas utiliser la variable de poids de l'enquête peut entraîner des estimations biaisées qui ne seront pas compatibles avec celles produites par Statistique Canada.

7.2 Variables de poids bootstrap

La variance exacte ne peut pas être calculée car cela nécessiterait une estimation pour chaque échantillon admissible alors qu'en pratique, les données ne sont collectées que pour un seul de ces échantillons, celui sélectionné aléatoirement par l'enquête. Cependant, cette variance peut être estimée en utilisant les variables de poids bootstrap fournies avec le fichier analytique.

Une procédure bootstrap adaptée au plan d'échantillonnage de l'EPLOSM a produit 1 000 variables de poids bootstrap, bsw1-bsw1000, qui s'utilisent afin d'estimer la variance d'une estimation. En principe, il s'agit de remplacer à tour de rôle dans l'analyse le poids d'enquête par les poids bootstrap afin d'obtenir 1 000 estimations dont la dispersion autour de l'estimation d'enquête constituera une estimation de la variance associée à cette dernière. En pratique, l'analyste peut simplement obtenir une telle estimation bootstrap de la variance en recourant à un logiciel statistique approprié – voir la section 8.6.1 pour de plus amples détails.

Important :

- Même les utilisateurs possédant une expertise de la méthodologie bootstrap doivent utiliser l'ensemble de 1 000 poids bootstrap de l'EPLOSM plutôt que d'entreprendre par eux-mêmes la mise en œuvre d'une approche bootstrap. La raison est que le fichier analytique ne contient pas toutes les informations d'enquête nécessaires afin de saisir adéquatement les diverses composantes d'enquête qui contribuent à la variance.
- En raison de la procédure bootstrap particulière employée pour l'EPLOSM, un facteur de dilatation de la variance égal à 16 doit être appliqué à la variance bootstrap calculée afin qu'elle reflète correctement la variance – voir la section 8.6 pour de plus amples détails.
- Sans égard à ce que la documentation de certains logiciels puisse laisser à entendre, il ne suffit pas d'indiquer au logiciel que le plan d'échantillonnage de l'EPLOSM afin d'en tirer des estimations de la variance appropriées : les variables de poids bootstrap doivent être utilisées. Il s'ensuit qu'il ne suffit pas de fournir au logiciel le poids d'enquête via son énoncé WEIGHT en le laissant estimer la variance sans inclure les variables de poids bootstrap fournies – voir la section 8.6 pour de plus amples détails.

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 diffusent d'une autre façon des données calculées à partir des fichiers 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.

L'annexe E propose un survol de l'analyse de données d'enquête complexe.

8.1 Lignes directrices concernant la confidentialité

La Loi interdit à Statistique Canada de divulguer toute information que l'organisme recueille qui pourrait permettre d'identifier une personne, une entreprise ou une organisation, sauf si le répondant a donné son consentement ou si la Loi sur la statistique le lui permet. Diverses règles de protection de la confidentialité sont utilisées afin de

prévenir la divulgation d'informations confidentielles. Lorsque cela est nécessaire, des résultats sont retirés afin d'éviter la divulgation directe ou résiduelle d'informations identifiables.

Des règles de protection de la confidentialité s'appliquent à toutes les données diffusées ou publiées, sans égard au mode d'accès aux données employé. Elles s'appliquent tout autant au personnel de Statistique Canada qu'aux personnes réputées être employées par l'agence qui utilisent les Centres de données de recherche (CDR). Ces règles sont conçues afin d'empêcher la publication ou la divulgation de toute information jugée confidentielle.

Dans le cas des chercheurs dans les CDR, deux niveaux de contrôle de la confidentialité sont prévus avant la diffusion d'estimations fondées sur l'EPLOSM. D'abord, tout chercheur ayant accès au fichier analytique confidentiel doit se conformer aux lignes directrices relatives à la diffusion qui sont énoncées dans le présent guide de l'utilisateur. Ensuite, tous les résultats sont vérifiés par un analyste désigné du CDR. **Il convient de noter que des règles de contrôle supplémentaires doivent être appliquées aux variables de revenu.** La décision de permettre ou non la sortie de résultats d'un CDR reposera sur le respect des lignes directrices en matière de confidentialité. Une fois que les lignes directrices relatives à la diffusion et à la publication ont été respectées, la décision concernant la présentation de résultats dans une publication relève de la responsabilité du chercheur et des responsables du processus d'examen par les pairs.

Le tableau 8.1 présente un résumé des lignes directrices en matière de confidentialité qui s'appliquent à l'EPLOSM. Toutes les données doivent être diffusées sous forme agrégée. En règle générale, il est interdit de diffuser des données non pondérées. Les fréquences non pondérées qui sous-tendent des estimations pondérées doivent être de 10 ou plus. Toutes les estimations pondérées doivent être arrondies. **Des règles plus strictes s'appliquent pour les estimations relatives aux variables du revenu du Recensement de 2021 ou aux estimations relatives à des données géographiques détaillées se situant à un niveau inférieur à celui des régions de l'EPLOSM**. La section 8.2 fournit des informations supplémentaires à l'égard des tailles d'échantillon minimales requises.

Tableau 8.1
Lignes directrices sur la confidentialité pour l'Enquête sur la population de langue officielle en situation minoritaire de 2022

Critère	Ligne directrice de 2022	Remarque
1) Quelle est la fréquence non pondérée minimale requise?	10	Pour les estimations du revenu, il doit y avoir au moins 10 unités avec un revenu non nul. Voir la sous-section 8.2 ci-dessous pour obtenir plus de renseignements à ce sujet.
2) Est-il permis de diffuser des résultats descriptifs non pondérés?	NON – Cela est interdit.	Les résultats non-pondérés de nature descriptive (p. ex., totaux, proportions, ratios, moyennes, corrélations) ne sont pas permis. Les tailles d'échantillon associées au modèle (statistique N) peuvent être relâchées sous réserve d'une justification fournie par l'analyste principal. Voir aussi le critère (3) ci-dessous.
3) Peut-on diffuser à la fois des résultats descriptifs pondérés et des résultats descriptifs non pondérés relativement à cette enquête?	NON	En général, une autorisation sera accordée uniquement dans le cas où une publication nécessiterait le fait d'avoir des tableaux de fréquences avec et sans pondération (une lettre du rédacteur en chef doit alors être présentée).
4) Peut-on diffuser les résultats d'un modèle avec et sans pondération pour cette enquête?	OUI	…
5) Comment définit-on des « données géographiques détaillées » pour cette enquête?	Toute totalisation se situant au-dessous de la région linguistique est considérée comme étant « détaillée ». Les régions métropolitaines de recensement (RMR) de Toronto, Montréal et Vancouver constituent des exceptions à cet égard.	Pour une justification du concept sous-jacent, veuillez communiquer avec le CDR. Voir également les sections 8.2 et 8.3 ci-dessous.
6) Variables géographiques restreintes	Les renseignements relatifs à la base de sondage (p. ex. les listes de codes postaux ou de subdivisions de recensement [SDR]) qui se situent en dessous du niveau géographique détaillé pour une enquête sont considérés comme étant confidentiels et ne peuvent être diffusés.	Ce critère se rapporte aux enquêtes ayant un niveau géographique restreint en deçà duquel il est possible de déceler des renseignements provenant de la base de sondage. Ce critère est inclus dans la documentation de chaque enquête, mais ne s'applique pas à l'EPLOSM.

Tableau 8.1
Lignes directrices sur la confidentialité pour l'Enquête sur la population de langue officielle en situation minoritaire de 2022

Critère	Ligne directrice de 2022	Remarque
7) Est-ce que tous les résultats descriptifs pondérés doivent être arrondis? Dans l'affirmative, quelle est la base d'arrondissement?	OUI Base d'arrondissement : ·10 pour les chiffres et les totaux ·100 pour les totaux des revenus ·10 pour les moyennes ou les médianes de revenu ·1 (c.-à-d. au nombre entier le plus proche) pour les ratios comportant des variables de revenu. Le nombre de décimales pour les proportions est limité à trois (0,001 ou 0,1 % pour les pourcentages).	Les résultats descriptifs d'un niveau géographique détaillé doivent être pondérés et arrondis à la base suivante : ·50 pour les chiffres et les totaux ·100 pour les totaux des revenus ·10 pour les moyennes ou les médianes de revenu ·1 (c.-à-d. le nombre entier le plus proche) pour les ratios comportant des variables de revenu. Le nombre de décimales pour les proportions à des niveaux géographiques détaillés est limité à deux (0,01 ou 1 % pour les pourcentages). Voir aussi la section 8.3 ci-dessous.
8) Quelles règles de confidentialité devraient être utilisées pour les estimations relatives aux variables du revenu du Recensement de 2021 qui ont été couplées avec l'EPLOSM?	Les CDR doivent se référer au document sur les règles de confidentialité du Recensement de 2021 qui doivent être appliquées en plus de celles de l'EPLOSM. Les services à la clientèle devraient communiquer avec l'équipe d'enquête pour obtenir les documents internes sur la règle sur le revenu. Ils devraient consulter ces documents en plus des règles de confidentialité mentionnées dans le présent guide.	Voir aussi les 8.2 et 8.3 ci-dessous.

... n'ayant pas lieu de figurer

Note : Pour les définitions de la géographie des RMR et de l'EPLOSM, voir le Dictionnaire, Recensement de la population, 2021.

Source : Enquête sur la population de langue officielle en situation minoritaire de 2022.

8.2 Lignes directrices sur la fréquence non pondérée minimale

Dans le contexte de l'EPLOSM de 2022, il importe de respecter une fréquence non pondérée minimale de manière à se conformer aux exigences en matière de confidentialité de la Loi sur la statistique. De plus, la fréquence non pondérée minimale constitue également une condition essentielle par rapport à la fiabilité des estimations. Voici les exigences minimales applicables aux fréquences non pondérées de toute statistique descriptive rattachée à l'EPLOSM.

- La fréquence non pondérée minimale est de 10. Toute estimation basée sur moins de 10 répondants doit être supprimée pour des raisons de confidentialité.
- Toutes les cellules des tableaux croisés qui ne respectent pas le critère minimal doivent être supprimées elles aussi pour des raisons de confidentialité.
- Tous les autres types de statistiques descriptives doivent être calculés en se fondant sur ce nombre minimal d'observations. Si la statistique descriptive est bivariée, le nombre d'observations doit être au moins égal à ce minimum pour chacune des deux variables contributrices. Par exemple, si un ratio est calculé, le numérateur et le dénominateur doivent tous deux être fondés sur le nombre minimal requis d'observations.
- Pour les estimations de revenus (p. ex. médianes, moyennes, totaux), il doit y avoir au moins 10 unités avec un revenu non nul dans une cellule donnée pour que la cellule soit publiée.

8.3 Lignes directrices pour l'arrondissement d'estimations

De manière à assurer que les estimations produites à partir du fichier analytique de l'EPLOSM concordent avec celles produites par Statistique Canada, il est fortement conseillé aux utilisateurs de se conformer aux lignes directrices suivantes en matière d'arrondissement. La diffusion d'estimations non arrondies pourrait induire les utilisateurs en erreur, car ces estimations pourraient donner l'impression d'être plus précises qu'elles ne le sont en réalité. De plus, l'arrondissement est une mesure de protection de la confidentialité qui est requise pour toutes les statistiques pondérées.

1. Les estimations de totaux figurant dans le corps d'un tableau statistique doivent être arrondies à la dizaine près au moyen de la méthode d'arrondissement classique.
2. Les totaux partiels et généraux des tableaux statistiques doivent être calculés à partir de leurs éléments non arrondis, puis être arrondis à la dizaine près au moyen de la méthode d'arrondissement classique.
3. Les moyennes, les proportions, les taux, les pourcentages et autres formes de ratios doivent être calculés à partir d'éléments non arrondis, ensuite être arrondis à la décimale près au moyen de la méthode d'arrondissement classique. Par exemple, une valeur de 0,1234 ou 12,34 % devrait être arrondie à 0,123 ou 12,3 %. Dans le cas des données géographiques détaillées (voir le tableau 8.1), les moyennes, les proportions, les taux, les pourcentages et autres formes de ratios doivent être arrondis au nombre entier le plus près selon la méthode d'arrondissement classique. Par exemple, une valeur de 0,1234 ou 12,34 % devrait être arrondie à 0,12 ou 12 %.
4. Les sommes et les différences d'agrégats ou de ratios doivent être calculées à partir de leurs éléments non arrondis, puis être arrondies à la dizaine ou à la décimale près au moyen de la méthode d'arrondissement classique.
5. Peu importe le niveau géographique, les totaux pondérés des revenus doivent être arrondis à la centaine près et les moyennes ou médianes pondérées des revenus doivent être arrondies à la dizaine près. Les ratios pondérés qui comprennent des variables de revenu doivent être arrondis au nombre entier le plus proche.
6. Les intervalles de confiance (IC) des estimations doivent être calculés à partir de leurs éléments non arrondis, puis arrondis à la dizaine près ou à la décimale près au moyen de la méthode d'arrondissement classique. Étant donné que l'estimation et les limites de confiance correspondantes sont arrondies de façon indépendante, l'estimation ne se situera pas toujours exactement au milieu de l'IC.
7. Dans le cas des données géographiques détaillées, toutes les statistiques doivent être arrondies à la cinquantaine près. En ce qui a trait à l'EPLOSM, toute totalisation se situant en dessous de la région linguistique est considérée comme étant « détaillée ». Les RMR de Toronto, Montréal et Vancouver constituent des exceptions à cet égard. À ce niveau de détail, il est important d'arrondir toutes les formes de ratios au nombre entier le plus près (voir le point 3 ci-dessus).
8. Il peut arriver, en raison de contraintes d'ordre technique ou autre, qu'une méthode d'arrondissement autre que la méthode classique soit utilisée. Dans un tel cas, les estimations obtenues pourraient ne pas correspondre à celles produites par Statistique Canada. Le cas échéant, il est fortement conseillé à l'utilisateur d'exposer la raison de ces divergences dans le document publié.

8.3.1 Méthode d'arrondissement classique

Selon la méthode d'arrondissement classique, si le premier ou le seul chiffre à supprimer est de 0 à 4 (par exemple le chiffre « 3 » dans le nombre « 823 » si l'arrondissement est fait à la dizaine près, ou le chiffre « 2 » si l'arrondissement est fait à la centaine près), le dernier chiffre retenu ne change pas (ainsi, dans le nombre « 823 », le chiffre « 2 » ne change pas si l'arrondissement est fait à la dizaine près, ce qui donnera « 820 », ou le chiffre « 8 » ne change pas si l'arrondissement est fait à la centaine près, ce qui donne « 800 »). Si le premier ou le seul chiffre à supprimer est de 5 à 9 (par exemple le chiffre « 5 » dans le nombre « 865 » si l'arrondissement est fait à la dizaine près, ou le chiffre « 6 » si l'arrondissement est fait à la centaine près), la valeur du dernier chiffre retenu augmente d'une unité (par exemple dans le nombre « 865 », le chiffre « 6 » augmente d'une unité lorsque l'arrondissement est fait à la dizaine près, ce qui donne « 870 », ou le chiffre « 8 » augmente d'une unité si l'arrondissement est fait à la centaine près, ce qui donne « 900 »).

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

L'EPLOSM utilise un plan d'échantillonnage et une méthode d'estimation complexe; les poids d'enquête ne sont donc pas égaux pour toutes les unités échantillonnées. Lorsqu'ils produisent des estimations et des tableaux statistiques, les utilisateurs doivent appliquer le poids d'enquête. Si le poids d'enquête n'est pas utilisé, alors les estimations calculées à partir des fichiers 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.

Tous les répondants dont les données figurent dans le fichier analytique ont satisfait à une norme de réponse préétablie pour l'enquête. Cependant, certaines données manquantes subsistent en raison de non-réponses à certains éléments du questionnaire (c'est-à-dire, la non-réponse partielle). Il s'agit de situations où les répondants n'ont pas répondu à une question particulière qui s'appliquait à eux.

La catégorie « Saut valide » n'est généralement pas considérée comme une valeur manquante car elle indique une question qui n'était pas destinée à ce répondant particulier. Il en va de même pour les catégories « Ne s'applique pas » des variables de recensement, qui sont équivalentes aux sauts valides. Le terme « Saut valide » n'est pas utilisé dans le recensement et, par conséquent, les variables de recensement dans le fichier analytique de l'enquête conservent les mêmes étiquettes de catégorie qui leur ont été attribuées pour le recensement.

L'inclusion ou l'exclusion de l'un ou l'autre des types de valeurs manquantes susmentionnés dans une tabulation dépend de l'objectif de l'analyse. Les utilisateurs devront définir leur domaine d'estimation (population totale d'intérêt) pour chaque variable, en tenant compte des valeurs manquantes pour chacune. Par exemple, ils devront évaluer la pertinence d'inclure les valeurs manquantes dans le dénominateur utilisé pour calculer les pourcentages. Dans certains cas, les chercheurs peuvent décider que les valeurs manquantes sont significatives pour leur question de recherche. Le fait qu'un répondant choisisse de ne pas répondre à une question peut constituer en soi une information utile.

8.5 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 (IC) 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

L'estimation et l'IC peuvent être diffusés sans mise en garde. Les utilisateurs de données devraient utiliser l'IC de 95 % pour déterminer si la qualité de l'estimation est suffisante. Veuillez noter que le « A » n'est pas un indicateur de qualité; il ne devrait pas être diffusé avec l'estimation.

Catégorie E

L'estimation et l'IC doivent porter la mention « E » (ou un identificateur semblable) et être accompagnés d'une mise en garde sur la qualité, qui indique qu'il faut utiliser les estimations avec prudence. Les utilisateurs de données devraient utiliser l'IC de 95 % pour déterminer si la qualité de l'estimation est suffisante.

Catégorie F

L'estimation et l'IC ne sont pas recommandés pour la diffusion. Ils sont jugés de si piètre qualité qu'ils ne se prêtent à aucune utilisation statistique : ils sont très instables, ce qui les rend peu fiables et potentiellement trompeurs. Si les analystes insistent pour diffuser des estimations de piètre qualité, même après avoir été informés de leur inexactitude, ces estimations doivent être accompagnées d'un avis de non-responsabilité de la part de Statistique Canada. Les analystes doivent tenir compte des mises en garde reçues et s'engager à ne pas diffuser, présenter, ni rapporter les estimations, directement ou indirectement, sans cet avis de non-responsabilité. Ces estimations doivent être signalées par la lettre F (ou un identificateur semblable) et la mise en garde suivante doit accompagner les estimations et les IC : « 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 tirées de ces données ne seront pas fiables et peuvent être invalides ».

Les règles qui permettent d'assigner une estimation à une catégorie de diffusion varient selon le type d'estimation.

8.5.1 Règles de diffusion pour les estimations

Les estimations de proportions et de chiffres sont dérivées à partir de variables binaires. Les estimations de chiffres sont des estimations du nombre total de personnes avec une caractéristique d'intérêt; autrement dit, elles représentent une somme pondérée d'une variable binaire (p. ex. nombre estimé d'adultes de langue officielle en situation minoritaire). Les estimations de proportions sont des estimations de proportions de personnes avec une caractéristique d'intérêt (p. ex. proportion estimée d'adultes de langue officielle en situation minoritaire dans la population canadienne). Les estimations de chiffres et de proportions peuvent aussi être calculées à partir de variables catégoriques, c'est-à-dire l'estimation du nombre ou de la proportion des personnes qui appartiennent à une catégorie.

Les règles de diffusion pour les proportions estimées sont fondées sur la taille de l'échantillon et la longueur de l'IC exprimée en points de pourcentage (p.p.). Les règles de diffusion pour d'autres statistiques, comme les chiffres estimés, sont fondées sur la taille de l'échantillon et la longueur relative de l'IC. Le tableau 8.2 présente les règles de diffusion pour l'EPLOSM, lorsque le domaine d'intérêt est soit à l'échelle nationale (tout le Canada), infranational à l'exception des capitales territoriales, ou composé de capitales territoriales.

Tableau 8.2
Les règles de diffusion pour les estimations de l'EPLOSM

Catégorie de diffusion	Règle		Mesure
	Pour les proportions[2]	Pour d'autres statistiques[3]	
Règles pour les estimations à l'échelle nationale			
A1 Peuvent être diffusées	"n ≥ 180 et Longueur de l'IC ≤ 14 p.p."	"n ≥ 180 et Longueur relative de l'IC ≤ 1.4"	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
E Peuvent être diffusées avec une mise en garde	"90 ≤ n < 180 ou Longueur de l'IC > 14 p.p."	"90 ≤ n < 180 ou Longueur relative de l'IC > 1.4"	Diffuser avec une mise en garde sur la qualité (lettre E); les utilisateurs devraient utiliser l'IC comme indicateur de qualité
F Ne peuvent être diffusées	n < 90 (peu importe la longueur de l'IC)	n < 90 (peu importe la longueur relative de l'IC)	Supprimer l'estimation et son IC pour des raisons de qualité (lettre F)
Règles pour les estimations à l'échelle infranationale			
A1 Peuvent être diffusées	"n ≥ 150 et Longueur de l'IC ≤ 14 p.p."	"n ≥ 150 et Longueur relative de l'IC ≤ 1.4"	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
E Peuvent être diffusées avec une mise en garde	"75 ≤ n < 150 ou Longueur de l'IC > 14 p.p."	"75 ≤ n < 150 ou Longueur relative de l'IC > 1.4"	Diffuser avec une mise en garde sur la qualité (lettre E); les utilisateurs devraient utiliser l'IC comme indicateur de qualité
F Ne peuvent être diffusées	n < 75 (peu importe la longueur de l'IC)	n < 75 (peu importe la longueur relative de l'IC)	Supprimer l'estimation et son IC pour des raisons de qualité (lettre F)
Règles pour les estimations à l'échelle des capitales territoriales			
A1 Peuvent être diffusées	"n ≥ 60 et Longueur de l'IC ≤ 14 p.p."	"n ≥ 60 et Longueur relative de l'IC ≤ 1.4"	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
E Peuvent être diffusées avec une mise en garde	"30 ≤ n <60 ou Longueur de l'IC > 14 p.p."	"30 ≤ n < 60 ou Longueur relative de l'IC > 1.4"	Diffuser avec une mise en garde sur la qualité (lettre E); les utilisateurs devraient utiliser l'IC comme indicateur de qualité
F Ne peuvent être diffusées	n < 30 (peu importe la longueur de l'IC)	n < 30 (peu importe la longueur relative de l'IC)	Supprimer l'estimation et son IC pour des raisons de qualité (lettre F)

1. À noter que « A » ne constitue pas un indicateur de qualité et il ne devrait pas être diffusé avec l'estimation; l'IC de 95 % est l'indicateur de qualité.

2. À noter que pour des estimations de proportions, n est défini comme le chiffre non pondéré du nombre de répondants au dénominateur (et non au numérateur) de la proportion.

3. À noter que pour des estimations de moyennes, n est défini comme le chiffre non pondéré du nombre de répondants contribuant à l'estimation y compris les valeurs de zéro. Pour des estimations de totaux et de chiffres, n est défini comme le chiffre non pondéré du nombre de répondants ayant des valeurs non nulles contribuant à l'estimation.

Source : Enquête sur la population de langue officielle en situation minoritaire de 2022.

8.5.2 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.

8.5.3 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 de calage.)
- 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.

8.6 Lignes directrices pour l'analyse statistique, l'estimation de la variance et la construction des 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'EPLOSM 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 de poids bootstrap a été généré. Les utilisateurs obtiennent de ces poids 1 000 estimations bootstrap en mimant chaque fois la façon que l'estimation d'enquête a été obtenue des données et du poids d'enquête. La variance que montrent ces estimations bootstrap constitue une estimation de la variance que présenteraient les estimations d'enquête si tous les échantillons admis par le plan de l'EPLOSM pouvaient être réalisés.

8.6.1 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 valides 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. Voir Gagné et al. (2014) pour de plus amples détails.

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. C'est pourquoi il est possible de voir l'option varmethod-brr dans des instructions SAS alors que les poids bootstrap fournis avec les données ont bel et bien été utilisés ; le recours à cette astuce n'est plus nécessaire dans les versions récentes de SAS puisque l'option varmethod=bootstrap est maintenant disponible.

Important : bien que certains progiciels offrent des procédures qui prévoient l'emploi d'une variable de poids en combinaison avec les données, p. ex., l'énoncé WEIGHT de PROC FREQ dans SAS, cela ne signifie pas que prendre en compte la variable de poids d'enquête dans l'analyse statistique de cette façon suffise à obtenir des variances valides. Pour illustrer, considérons un échantillon aléatoire simple sans remise de taille 5 avec un poids d'enquête de 2 ; cela signifie donc que la population compte 5x2=10 unités. Cependant, si les données d'enquête échantillonnées étaient soumises à PROC FREQ avec le poids d'enquête de 2 dans son énoncé WEIGHT, SAS en déduirait les éléments suivants :

- la taille de l'échantillon est de 10, interprétant le poids d'enquête comme de simples fréquences, c'est-à-dire en voyant l'échantillon comme consistant en deux copies identiques du fichier contenant 5 unités qui a été spécifié comme entrée pour la procédure ;
- la taille de la population est infinie ;
- le plan d'échantillonnage est semblable à un échantillonnage aléatoire simple avec remise.

Ces hypothèses tacitement faites par le logiciel produiront une évaluation statistique erronée de l'incertitude dans ce contexte d'enquête. Une analyse appropriée des données d'enquête utiliserait plutôt PROC SURVEYFREQ de la suite des procédures d'enquête SAS, qui permet d'utiliser les variables de poids Bootstrap fournies avec les données (voir la documentation SAS concernant l'instruction REPWEIGHTS en conjonction avec l'option varmethod=Bootstrap).

8.6.2 Facteur d'ajustement de Fay ou de dilatation de la variance

Tel que mentionné à la section 7.2, il est très important d'appliquer un facteur multiplicatif d'accroissement C de 16 à la variance bootstrap standard calculée à partir des données afin d'obtenir une estimation de la variance qui soit valide dans le cadre de l'EPLOSM. La plupart des progiciels permettent à l'utilisateur d'indiquer la valeur à utiliser par l'entremise de ce qui est souvent appelé le facteur d'ajustement de Fay ou le coefficient de Fay.

Mise en garde : puisque les conventions diffèrent entre progiciels, il est recommandé de consulter la documentation du progiciel employé afin de déterminer la valeur à indiquer pour le facteur de Fay. Par exemple, dans SUDAAN la valeur à indiquer est directement C alors qu'en SAS le facteur de Fay doit valoir $1 - 1/\sqrt{C}$ afin que la variance calculée à la base soit multipliée par C tel qu'exigé. Ainsi, en SAS il faut indiquer une valeur de 0,75 afin d'obtenir le facteur de dilatation de la variance requis de 16.

8.6.3 Intervalles de confiance

Les intervalles de confiances les plus couramment utilisés, sont ceux de Wald et de Student ; ce sont souvent les méthodes par défaut des logiciels statistiques[9]. En ce qui concerne les proportions, on sait que l'hypothèse de normalité sous-jacente à ces méthodes ne tient pas dans le cas d'un échantillon 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; Liu et Kott, 2009; Neusy et Mantel, 2016).

9. Les intervalles de confiance de Wald et Student sont présentés dans l'annexe D, ainsi que dans les sections 7.2 et 7.3 du Rapport technique sur l'échantillonnage et la pondération, Recensement de la population, 2021.

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.

1. Anciennes versions de SAS, intervalles de confiance de Wilson modifiés :

   PROC SURVEYFREQ

   DATA=…. VARMETHOD=BRR (**FAY**=0.75);

   WEIGHT weight;

   REPWEIGHTS bsw1-bsw1000;

   TABLES …. / **CL (TYPE=WILSON ADJUST=NO TRUNCATE=YES)**

2. Versions récentes de SAS, intervalles de confiance de Wilson modifiés :

   PROC SURVEYFREQ

   DATA=…. VARMETHOD=bootstrap;

   WEIGHT weight;

   REPWEIGHTS bsw1-bsw1000 / **REPCOEFS**=0.016;

   TABLES …. / **CL (TYPE=WILSON ADJUST=NO TRUNCATE=YES)**

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

   PROC CROSSTAB

   DATA=…. DESIGN=BRR **SMCONF**=50;

   WEIGHT weight;

   REPWGT bsw1-bsw1000 / **ADJFAY**=16;

   TABLES ...;

Mise en garde : les poids bootstrap doivent être fournis au logiciel statistique employé, car le fait de spécifier uniquement la méthode d'estimation de la variance (via les instructions VARMETHOD pour SAS ou DESIGN pour SUDAAN) entraînera des résultats incorrects. En effet, en l'absence de poids bootstrap, la plupart des logiciels tenteront de mettre en œuvre eux-mêmes la procédure bootstrap. Cependant, il n'existe aucun logiciel statistique capable d'exécuter correctement la méthode Bootstrap de manière autonome dans le contexte de l'EPLOSM, même entre les mains de l'utilisateur le plus expérimenté, car une dérivation correcte des poids bootstrap de l'enquête nécessite des informations qui ne sont pas fournies dans le fichier analytique.

8.6.4 Rééchelonnement du poids d'enquête

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 du poids d'enquête, mais pas les poids bootstrap.

Cependant, ces procédures peuvent différer dans leur définition d'un poids, et produire des estimations d'enquête appropriées mais des estimations de leur variance invalides. Dans l'exemple donné à la Section 7.0 d'un échantillon aléatoire simple de taille 100 sélectionné à partir d'une population de taille 5 000, une telle procédure interpréterait le poids de sondage de 50 d'une unité échantillonnée comme un poids de fréquence : il y a 50 unités de ce type dans l'échantillon. Ainsi, lorsqu'on présente à la procédure un fichier de données de 100 unités ayant chacune un poids d'enquête de 50, elle considérerait avoir un échantillon de 5 000 au total (c'est-à-dire 50 copies de l'échantillon de l'enquête) qui aurait été tiré d'une population de taille infinie. En plus d'ignorer que le plan d'échantillonnage de l'enquête est complexe (notamment parce que la population est de taille finie et donc juste une fraction plus grande que l'échantillon), cette approche surestime ce que les données de l'enquête permettent de conclure.

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. Le poids pour le domaine d'intérêt doit être rééchelonné de sorte que le poids moyen est d'un (1); ceci peut être accompli en divisant chaque valeur de poids par le poids moyen global du domaine avant de procéder à l'analyse. Le rééchelonnement 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 lorsqu'aucune procédure permettant l'utilisation des poids bootstrap n'est disponible; les utilisateurs sont avisés que les résultats sont approximatifs.

8.6.5 Utilisation d'intervalles de confiance pour déterminer l'importance statistique d'un résultat

Tel que discuté en plus amples détails dans l'annexe E, deux situations se présentent fréquemment en pratique :

1. 1) établir si l'estimation obtenue appuie ou pas la thèse que la vraie valeur soit celle postulée, p. ex. « est-ce que la vraie proportion des femmes qui présentent une certaine caractéristique est 15 %? »
2. 2) établir si les vraies valeurs de deux groupes sont différentes ou pas, p. ex. « est-ce que la vraie proportion des femmes qui présentent une caractéristique diffère de celle des hommes? ».

Un intervalle de confiance peut être utilisé pour répondre à de telles questions. Dans le premier cas, il suffit de voir si la valeur postulée se trouve ou non dans l'intervalle de confiance associé à l'estimation. Par exemple, puisque 15 % se trouve dans l'IC [10 %, 23 %], on dirait alors qu'on ne peut pas rejeter sur la base des données l'hypothèse que la vraie proportion soit 15 %.

Lorsqu'on compare les estimations entre des groupes, la quantité d'intérêt devient la différence entre les deux estimations. Les IC peuvent être utilisés pour déterminer si la différence entre deux estimations est statistiquement significative ou non. Deux méthodes peuvent être utilisées pour déterminer l'importance statistique : comparer le chevauchement entre les IC des deux estimations et construire l'IC pour la différence des deux estimations. Ces approches sont discutées dans les deux prochaines sections.

8.6.6 Comparaison du chevauchement entre les intervalles de confiance des deux estimations

Lorsqu'on examine la différence de deux estimations afin de déterminer si les paramètres correspondants sont différents, une approche pratique consiste à comparer leurs IC afin de voir s'ils se chevauchent. On procède comme suit. Lorsque les deux IC se chevauchent, même très faiblement, il n'est pas possible de conclure quoi que ce soit à propos de la vraie différence entre les paramètres. Dans un tel cas, on ne peut pas rejeter l'hypothèse de travail que les deux paramètres soient égaux. Cependant, si les deux intervalles ne se chevauchent pas, alors on peut conclure que les paramètres de population concernés sont différents.

Cette méthode est conservatrice, au sens que les deux paramètres d'intérêt peuvent avoir des valeurs différentes et malgré tout produire des IC pour les estimations correspondantes qui se chevauchent. Autrement dit, la procédure n'est pas suffisamment fine afin de conclure à la différence des paramètres lorsque celle-ci existe bel et bien ; on dit dans le jargon statistique que ce test manque de puissance. Le caractère conservateur de l'approche s'explique par le fait qu'elle correspond à un test statistique qui utilise la plus grande variance possible pour la différence des estimations et non sa véritable variance. Par conséquent, afin que ce test soit concluant, il faut que la différence observée entre les estimations soit plus importante que la magnitude maximale de l'erreur. La

prochaine section décrit un test statistique qui repose sur la variance effective des estimations de la différence et qui est donc plus puissant.

8.6.7 Construction de l'intervalle de confiance pour la différence des deux estimations

L'autre possibilité est de calculer l'IC directement pour la différence des deux estimations: si cet IC contient zéro, alors on peut statistiquement conclure que les deux estimations ne diffèrent pas; si zéro ne se trouve pas dans l'IC, alors on rejette l'hypothèse d'égalité des paramètres. On peut notamment obtenir cet IC pour une différence par la méthode percentile-bootstrap décrite dans l'annexe D.

8.6.8 Diffusion d'informations statistiques à propos du genre

Dans le cadre des diffusions de l'EPLOSM de 2022, l'emploi de la variable sur le genre à deux catégories sera la norme dans les tableaux de données et les analyses. Les utilisateurs des données de l'EPLOSM sont encouragés à inclure les notes ci-dessous dans les tableaux ou les analyses qu'ils publient lorsqu'ils utilisent le genre.

Toutes les analyses et tous les tableaux de données diffusés selon le genre à deux catégories devraient comprendre les notes suivantes pour le genre, hommes+ et femmes+ :

Genre

Étant donné que la taille de la population non binaire est petite, il est nécessaire la plupart du temps d'agréger les données dans une variable sur le genre à deux catégories pour protéger la confidentialité des réponses. Dans ces cas, les personnes dans la catégorie « personnes non binaires » sont réparties dans les deux autres catégories de genre et sont désignées par le signe +.

Hommes+

Cette catégorie comprend les hommes et les garçons cisgenres et transgenres, de même que certaines personnes non binaires.

Femmes+

Cette catégorie comprend les femmes et les filles cisgenres et transgenres, de même que certaines personnes non binaires.

Dans les produits analytiques, il est préférable d'utiliser les termes « hommes/garçons » et « femmes/filles » dans le texte principal au lieu de « hommes+ » et de « femmes+ ». Toutefois, les tableaux, les graphiques et les autres images qui figurent dans les articles devraient utiliser les termes « hommes+ » et « femmes+ » et inclure les notes pertinentes.

Tous les tableaux de données et toutes les analyses comportant des comparaisons historiques sur la variable « sexe » en 2006 et la variable sur le genre à deux catégories utilisée à compter de 2022 devraient inclure les notes susmentionnées ainsi que la note suivante :

Les données sur le genre n'ont pas été recueillies dans le cadre de l'Enquête sur la vitalité des minorités de langue officielle (EVMLO) de 2006. Par conséquent, la variable sur le sexe de 2006 et la variable sur le genre à deux catégories de 2022 sont combinées dans cette analyse. Bien que le sexe et le genre soient deux concepts différents, l'introduction de la variable sur le genre ne devrait pas avoir d'incidence importante sur l'analyse des données et la comparabilité historique, étant donné la petite taille des populations transgenre et non binaire.

Cette note supplémentaire pourrait également être ajoutée :

Bien que le sexe et le genre renvoient à deux concepts différents, la terminologie relative au genre est utilisée dans ce tableau de données pour en faciliter la lecture.

8.6.9 Combiner les échantillons d'enfants et d'adultes

La combinaison des données recueillies par l'EPLOSM à propos des enfants et des adultes nécessite une grande prudence pour pouvoir tirer des conclusions statistiques valides et pertinentes concernant une même population à l'étude. D'abord, il faut décrire et restreindre avec précision la population étudiée. Par exemple, il serait incorrect de parler « des enfants et des adultes de langue officielle en situation minoritaire » puisque l'inclusion d'un enfant dans l'enquête est fondée sur le profil linguistique de ses parents et non du sien. Dans un tel cas, on décrira plutôt la population visée comme étant celle « des adultes de langue officielle en situation minoritaire et de leurs enfants ». Afin de rassembler les données pertinentes à cette population, il est nécessaire de ne conserver que les enfants qui avaient au moins un adulte de langue officielle en situation minoritaire dans leur famille de recensement en 2021 (LOSM_C_P = 1) lors de la mise en commun des fichiers d'adultes et d'enfants.

Ensuite, il faut s'assurer que les données utilisées lors de l'analyse proviennent de questions d'enquête qui sont identiques, ou du moins suffisamment similaires, pour les deux segments, enfants et adultes. En particulier, de telles questions doivent porter directement sur l'enfant, et non sur l'adulte qui répond à l'enquête en son nom sinon la mise en commun des échantillons d'enfants et d'adultes entraînera une surreprésentation des adultes et des estimations erronées pour cette population.

Il est important de noter qu'en combinant les fichiers d'adultes et d'enfants on n'obtient pas un échantillon d'individus qui soient apparentés. D'une part on retrouve parmi les adultes de langue officielle en situation minoritaire échantillonnés des personnes qui n'ont pas d'enfants, ou qui ont des enfants qui n'ont pas été échantillonnés. D'autre part, bien que les données concernant les enfants aient été collectées par l'entremise d'adultes répondants, ces adultes apparentés peuvent très bien ne pas avoir été échantillonnés, et même s'ils l'avaient été, les données des fichiers rendus disponibles ne permettraient pas d'établir les liens familiaux entre les personnes des fichiers des adultes et des enfants.

8.7 Différences entre l'EPLOSM et d'autres sources de données

En raison de la mise à jour du questionnaire et de certaines différences entre la méthodologie utilisée pour l'EPLOSM de 2022, le Recensement de 2021 et l'EVMLO de 2006, il faut faire preuve de prudence lorsque l'on compare les données entre ces sources. Les sections qui suivent fournissent des renseignements sur les éléments ayant une incidence sur la comparabilité des données, fournissant ainsi aux utilisateurs de données des renseignements importants sur les facteurs à considérer au moment d'effectuer des analyses à l'aide des données de l'EPLOSM de 2022 et de l'EVMLO de 2006.

L'échantillon de l'EPLOSM de 2022 a été sélectionné à partir des répondants ayant fourni des réponses précises au Recensement. Des renseignements plus détaillés sur la façon dont les réponses du Recensement ont été utilisées pour déterminer la population cible de l'EPLOSM sont fournis à la section 3 (méthodologie de l'enquête).

8.7.1 Différences entre l'EPLOSM de 2022 et le Recensement de 2021

Un des objectifs de l'EPLOSM était de contribuer à répondre aux besoins en matière de données qui ne sont pas comblés par les données existantes, tel que le Recensement. Le Recensement et l'EPLOSM sont deux sources importantes d'information sur la population de langue officielle en situation minoritaire qui se complètent l'une l'autre. L'EPLOSM repose sur des concepts qui sont abordés dans le Recensement et comporte des questions plus approfondies pour produire des renseignements plus détaillés. Par exemple, les questions 8 à 10 du Recensement fournissent des renseignements sur la connaissance et l'usage des langues et les questions 12 à 17 sur la langue d'instruction aux niveaux primaire et secondaire. L'ajout de renseignements provenant de l'EPLOSM offre l'occasion d'en apprendre davantage sur les compétences linguistiques des adultes et des enfants ainsi que sur la langue d'enseignement des études postsecondaires des adultes et à différents niveaux scolaires pour les enfants, par exemple, les intentions des parents quant à la langue d'instruction de leurs enfants et leurs raisons et préférences en matière de choix de langue d'instruction de leurs enfants.

L'EPLOSM porte aussi sur des sujets ou des thèmes qui ne sont pas couverts par le Recensement. Par exemple, l'EPLOSM fournit des renseignements sur le fait qu'un enfant en situation minoritaire ait fréquenté une garderie dans

les langues officielles et sur l'accès ou l'utilisation des services dans la langue officielle minoritaire par les adultes de langue officielle en situation minoritaire. L'EPLOSM donne l'occasion d'obtenir davantage de renseignements détaillés auprès des enfants ou des adultes sur leur expérience en matière de dynamiques linguistiques, d'arts, de culture et de médias, de sentiment d'appartenance, de vitalité, de participation communautaire et d'utilisation des langues officielles dans l'espace public.

Les comptes de population tirés de l'EPLOSM de 2022 pour certaines sous-populations peuvent différer de ceux obtenus à partir du Recensement, et ce, même lorsque l'univers de la population du Recensement est restreint à celui de l'EPLOSM. La pondération fait en sorte que le nombre d'adultes de la population de langue officielle en situation minoritaire soit le même dans le Recensement et l'EPLOSM, mais seulement pour certaines combinaisons de cette population, de région et de groupe d'âge des strates de l'échantillon. De même, le nombre d'enfants sera identique qu'il soit calculé à partir du Recensement ou de l'EPLOSM, mais seulement pour certaines combinaisons des régions et groupe d'âge. Toutefois, les comptes de population peuvent différer pour d'autres sous-populations, qui n'ont pas été contrôlées lors de l'échantillonnage et la post-stratification (voir sections 3 et 7).

Par ailleurs, bien que certains concepts soient communs à l'EPLOSM et au Recensement de la population de 2021, les résultats peuvent différer légèrement entre ces deux sources. Pour une même personne, les caractéristiques déclarées peuvent, dans certains cas, différer entre l'EPLOSM et le Recensement. Il existe de nombreuses raisons pour lesquelles les réponses provenant de ces sources de données peuvent différer, notamment, l'effet du temps et de la déclaration par personne interposée.

L'effet du temps qui s'est écoulé entre le Recensement de 2021 et la période de collecte de l'EPLOSM en 2022 peut expliquer que certaines caractéristiques peuvent varier, en particulier pour les caractéristiques qui ne sont pas fixes dans le temps.

Dans la plupart des régions, les données du Recensement de 2021 ont été recueillies par autodénombrement. Les questionnaires ont été remplis en ligne ou sur papier et retournés par la poste.

Souvent, une même personne a rempli le questionnaire du Recensement pour tous les membres du ménage. On parle alors de déclaration par personne interposée. Les données de l'EPLOSM ont été recueillies, dans la plupart des cas, par autodéclaration au moyen d'un questionnaire électronique auprès de l'adulte sélectionné ou d'un parent de l'enfant sélectionné. Puisque la personne contactée dans le cadre de l'EPLOSM n'est pas nécessairement celle qui a rempli le questionnaire du Recensement, les réponses à des questions similaires peuvent varier.

8.7.2 Comparabilité entre l'EPLOSM de 2022 et l'EVMLO de 2006

L'Enquête sur la population de langue officielle en situation minoritaire (EPLSOM) a été conçue afin de fournir des données sur les enjeux et les sujets actuels qui ont une incidence sur les minorités de langue officielle tout en permettant, dans la mesure du possible, de mesurer l'évolution de la situation des communautés de langue officielle en situation minoritaire depuis la dernière enquête similaire, soit l'Enquête sur la vitalité des minorités de langue officielle (EVMLO) de 2006.

Les principaux nouveaux sujets abordés dans le cadre de l'EPLOSM de 2022 sont le niveau de scolarité des parents des adultes et des enfants, les intentions des parents quant à la langue d'instruction de leurs enfants, l'hésitation des adultes à utiliser la langue officielle minoritaire, la perception de discrimination pour des raisons linguistiques chez les parents et les adultes, les services d'aide à l'intégration des personnes immigrantes, et l'impact de la pandémie de COVID-19 sur l'accès et l'utilisation des services dans la langue officielle minoritaire. Il n'est donc pas possible d'effectuer de comparaison avec la situation qui prévalait en 2006 pour ces nouveaux sujets.

De plus, le questionnaire de l'EVMLO de 2006 a été actualisé afin d'assurer la pertinence de l'EPLOSM de 2022. Certaines questions de l'EPLOSM de 2022 sont restées identiques ou suffisamment similaires à celles de l'EVMLO de 2006 pour assurer la comparabilité de plusieurs résultats entre ces deux sources (voir les questionnaires et les dictionnaires des codes des deux enquêtes). Cependant, certains de ces changements limitent la comparabilité de certains résultats de l'EPLOSM de 2022 avec ceux de l'EVMLO de 2006. Par exemple, le concept de sexe utilisé dans l'EVMLO de 2006 a été remplacé par celui de genre dans l'EPLOSM de 2022. Dans de rares cas, des changements importants aux questions d'un module et à leur structure ne permettent pas la comparabilité dans le

temps comme c'est le cas du module portant sur les services de santé de l'EPLOSM de 2022. L'annexe A fournit des indications générales quant à la comparabilité avec l'EVMLO de 2006 des principaux concepts contenus dans les différents modules de l'EPLOSM de 2022. Une partie du contenu de l'EVMLO de 2006 n'a pas été repris dans l'EPLOSM de 2022 afin de limiter le fardeau de réponse des répondants. Toutefois, plusieurs renseignements disponibles dans le Recensement de 2021 ont été ajoutés aux fichiers de microdonnées de l'EPLOSM de 2022.

Outre les différences de contenu, les échantillons de l'EVMLO de 2006 et de l'EPLOSM de 2022 comportent aussi quelques différences dont il faut tenir compte lors de comparaisons entre ces deux sources. Les deux principales différences, décrites dans les deux paragraphes qui suivent, sont l'ajout en 2022 de certains enfants admissibles à l'instruction dans la langue officielle minoritaire dans le fichier des enfants et le retrait de l'échantillon supplémentaire de personnes de langue maternelle tierce ayant le français comme première langue officielle parlée dans le fichier des adultes.

L'échantillon d'enfants de l'EPLOSM de 2022 inclut des enfants admissibles à l'instruction dans la langue officielle minoritaire qui n'avaient pas de parent faisant partie de la population de langue officielle minoritaire. Les nouvelles questions sur la langue d'instruction ajoutées au Recensement de 2021 permettaient de les inclure, ce qui n'était pas possible en 2006. C'est pourquoi lorsqu'on compare les données de l'EPLOSM de 2022 sur les enfants, ces comparaisons ne doivent porter que sur les enfants qui avaient au moins un adulte de langue officielle en situation minoritaire dans leur famille de Recensement en 2021 (LOSM_C_P = 1), afin que cet univers corresponde à celui des enfants de l'EVMLO de 2006.

La principale différence entre les fichiers d'adultes de l'EVMLO de 2006 et de l'EPLOSM de 2022 concerne le fait que le fichier de l'EVMLO de 2006 inclut un échantillon d'adultes « allophones de première langue officielle parlée (PLOP) français » qui a été sondé dans la région de Montréal pour des besoins d'analyses spécifiques (voir la section 3.2 de ce guide et le guide de l'EVMLO). Comme cet échantillon ne faisait pas strictement partie de la population cible de 2006, ces adultes doivent être exclus du fichier de l'EVMLO de 2006 lors de comparaisons incluant les adultes de la région de Montréal.

D'autres différences touchent quelques régions géographiques. Comme l'indique le tableau synthèse de l'annexe A, il n'est pas possible de comparer les estimations pour les territoires issues de l'EVMLO de 2006 à celles de l'EPLOSM de 2022, car en 2022 seules les capitales territoriales étaient incluses dans l'échantillon (voir la section 3.1 de ce guide). Les autres différences entre les régions géographiques de l'EVMLO de 2006 et de l'EPLOSM de 2022 touchent les régions du Nouveau-Brunswick qui sont légèrement différentes dans l'EPLOSM de 2022. La région du Nord du Nouveau-Brunswick inclut maintenant les deux subdivisions de recensement (SDR) de 2021 à forte présence francophone d'Alnwick et de Neguac. La région du Sud-Est du Nouveau-Brunswick inclut aussi les subdivisions de recensement (SDR) de 2021 de Rogersville (paroisse et village) et de Hardwicke qui comprend la communauté francophone de Baie-Sainte-Anne. Les comparaisons avec les régions de cette province de l'EVMLO de 2006 doivent être faites avec prudence, en particulier pour la région du « reste du Nouveau-Brunswick ».

Les autres différences entre les échantillons de l'EVMLO de 2006 et de l'EPLOSM de 2022 sont mineures et ont généralement un effet négligeable sur les comparaisons d'ensemble entre ces deux sources. Par exemple, les deux enquêtes postcensitaires excluaient les résidents permanents de leurs échantillons; cependant la méthode d'identification des résidents permanents a changé depuis 2006.

Finalement, deux sous-groupes de langue de la population cible de l'EVMLO de 2006 n'ont pas été inclus dans l'échantillon de l'EPLOSM de 2022. Il s'agit des personnes n'ayant ni le français, ni l'anglais comme langue maternelle, connaissant les deux langues officielles et parlant les deux langues officielles également le plus souvent à la maison (seules ou avec d'autres langues), ainsi que des rares personnes ayant comme langues maternelles la langue officielle majoritaire et une langue tierce sans la langue officielle minoritaire et connaissant uniquement la langue officielle minoritaire. Bien que ces deux sous-groupes représentaient moins de 1 % de la population cible de l'EVMLO de 2006 (voir les tableaux B.1 et B.3 du guide méthodologique de l'EVMLO de 2006), il est néanmoins préférable de les retirer lors de comparaisons avec l'EPLOSM de 2022.

De plus, il est important de noter que le fichier des enfants de l'EVMLO de 2006 ne comporte pas d'identifiant unique et qu'en conséquence deux variables (sampleid et childid) doivent être utilisées afin d'ordonner les fichiers avant d'y lier les poids bootstrap.

Annexe A : Les thèmes et concepts du contenu de l'EPLOSM de 2022 et la comparabilité avec l'EVMLO de 2006

Tableau A
Les thèmes et concepts du contenu de l'EPLOSM de 2022 et la comparabilité avec l'EVMLO de 2006

Thèmes	Modules du questionnaire de 2022	Fichiers maitres	Principaux contenu et concepts	Comparabilité avec l'EVMLO de 2006
Sociodémo-graphique	Identification du répondant (NAM)	Adultes & Enfants	Age de l'adulte ou du parent de l'enfant en date du 16 mai 2022.	Groupes d'âge comparables : 18-24, 25-44, 45-64, 65+.
Sociodémo-graphique	Identification du répondant (NAM2)	Enfants	Age de l'enfant en date du 16 mai 2022.	Groupes d'âge comparables : 1-4, 5-11, 12-17.
Sociodémo-graphique	Identification du répondant (ANU)	Adultes & Enfants	Province de résidence en date du 16 mai 2022.	Comparables à celles de 2006, sauf les territoires.
Sociodémo-graphique	Genre du répondant ciblé (GDR)	Adultes & Enfants	Genre de l'adulte ou de l'enfant (voir les variables GND)	Nouveau en 2022. Le sexe a été utilisé en 2006.
Sociodémo-graphique	Identification du répondant (LAN)	Adultes	Principales caractéristiques linguistiques.	Comparables à celles de 2006.
Sociodémo-graphique	Composition du ménage (HHC)	Adultes & Enfants	Nombre de personnes, leur âge, genre et lien de parenté avec le répondant.	Le genre est nouveau en 2022 (voir variables GND).
Sociodémo-graphique	Composition du ménage (MS)	Adultes & Enfants	L'état matrimonial du répondant ou du parent de l'enfant et des conjoint(e)s.	Étendu à l'ensemble des membres du ménage en 2022.
Sociodémo-graphique	Conjoint(e) du répondant (SP)	Adultes & Enfants	Lien de parenté avec l'enfant, caractéristiques linguistiques et lieu de naissance.	Généralement comparable, une question retirée en 2022.
Sociodémo-graphique	Parents du répondant (PAR)	Adultes & Enfants	Lieu de naissance, langue et éducation du père et de la mère du répondant.	Comparable à 2006, sauf l'éducation des parents.
Sociodémo-graphique	Enfants présents dans le ménage (CHD)	Adultes & Enfants	Caractéristiques linguistiques de chaque enfant (< 18 ans) vivant dans le ménage.	Étendu à l'ensemble des enfants dans le ménage en 2022.
Langue	Compétences linguistiques (LSK)	Adultes & Enfants	Compétences linguistiques de l'adulte ou l'enfant. Nouveau en 2022: discrimination.	Module mis à jour. En partie comparable à 2006.
Langue	Compétences linguistiques du parent de l'enfant (LSP)	Enfants	Compétences linguistiques du parent de l'enfant. Nouveau en 2022: discrimination.	Module mis à jour. En partie comparable à 2006.
Éducation	Scolarité des adultes (EDU)	Adultes	Établissement postsecondaire fréquenté, lieu, langue et raisons de la langue.	Généralement comparable à 2006.
Éducation	Scolarité des enfants du répondant (EDC)	Enfants	Éducation de l'enfant : institution, langue, satisfaction des parents, raisons et préférences.	Module mis à jour. En partie comparable à 2006.
Éducation	Intentions quant à la scolarité (INT)	Enfants	Intentions éducatives des parents pour leur enfant et raisons de ces intentions.	Nouveau module en 2022. Non comparable à 2006.
Éducation	Services à la petite enfance (CHS)	Enfants	Langue(s) de la garderie de l'enfant, raisons et préférences des parents.	Module mis à jour. En partie comparable à 2006.
Langue	Trajectoire linguistique de l'enfance à l'âge adulte (TRA)	Adultes & Enfants	Trajectoires linguistiques de l'enfance à l'âge adulte de l'adulte ou du parent.	Module mis à jour. En partie comparable à 2006.
Langue	Dynamique linguistique familiale de l'enfant (DYN)	Enfants	Langue parlée par l'enfant à la maison, avec ses amis et à l'école.	Module mis à jour. En partie comparable à 2006.
Identité	Sentiment d'appartenance et vitalité perçue (VIT)	Adultes & Enfants	L'identité linguistique de l'adulte ou du parent et sa perception de la situation.	Module mis à jour. En partie comparable à 2006.
Participation	Participation communautaire (CIV)	Adultes	Participation des adultes à des activités bénévoles et à des associations.	Module mis à jour. En partie comparable à 2006.
Langue	Utilisation des langues dans l'espace public (PUB)	Adultes & Enfants	Utilisation des langues par les adultes, les enfants et les parents dans l'espace public.	En grande partie nouveau. Peu comparable à 2006.

Tableau A
Les thèmes et concepts du contenu de l'EPLOSM de 2022 et la comparabilité avec l'EVMLO de 2006

Thèmes	Modules du questionnaire de 2022	Fichiers maitres	Principaux contenu et concepts	Comparabilité avec l'EVMLO de 2006
Services	Gouvernement (GOV)	Adultes	Préférence, accès, usage, satisfaction à l'égard de la langue des services gouvernementaux.	Module mis à jour. En partie comparable à 2006.
Services	Justice (JUS)	Adultes	Accès, usage, satisfaction à l'égard de la langue des services de police et de justice.	Module mis à jour. En partie comparable à 2006.
Services	Santé (HLT)	Adultes	Importance, accès, usage, satisfaction à l'égard de la langue des services de santé.	Module mis à jour. Non comparable à 2006.
Services	Services aux immigrants (IMS)	Adultes	Langue des services obtenus à l'arrivée au Canada par les immigrants adultes.	Nouveau module en 2022. Non comparable à 2006.
Culture	Arts, culture et médias (ACM)	Adultes & Enfants	Langue des activités artistiques, culturelles, médiatiques de l'adulte ou de l'enfant.	Module mis à jour. En partie comparable à 2006.
Mobilité	Mobilité géographique (GEO)	Adultes & Enfants	Lieu de résidence à la naissance, à 15 ans et au moment de l'enquête. Raisons des migrations	Module mis à jour. En partie comparable à 2006.
Travail	Marché du travail (LAB)	Adultes & Enfants	L'activité professionnelle de l'adulte ou du parent, le type et le lieu de leur emploi.	Module mis à jour. En partie comparable à 2006.
COVID	COVID-19 (COV)	Adultes & Enfants	Variation d'usage des langues en public par l'adulte ou le parent pendant la COVID-19.	Nouveau module en 2022. Non comparable à 2006.

Référence : Enquête sur la vitalité des minorités de langue officielle (EVMLO), 2006
Source : Enquête sur la population de langue officielle en situation minoritaire de 2022

Annexe B. Exemples de calculs d'estimations et d'intervalles de confiance

Cette annexe fournit un exemple de calcul d'une estimation ponctuelle et de son intervalle de confiance (IC) à partir de l'Enquête sur la population de langue officielle en situation minoritaire (EPLOSM). Il est à noter que pour déterminer l'intervalle de confiance (IC) des estimations, il faut utiliser un logiciel ou un progiciel statistique qui peut calculer les estimations de variance à partir des poids bootstrap fournis avec les données. Dans les exemples suivants, c'est le logiciel SAS qui est utilisé.

Estimation du pourcentage d'adultes de langue officielle minoritaire ayant vécu une situation où ils ont hésité à utiliser la langue officielle minoritaire au cours des cinq années précédant l'enquête (voir la section 2.2 Insécurité linguistique dans le rapport analytique de l'enquête Situation des populations de langue anglaise au Québec et de langue française au Canada hors Québec : résultats de l'Enquête sur la population de langue officielle en situation minoritaire de 2022**).**

Commencez par consulter la section relative au module « compétences linguistiques » (LSK) dans l'index thématique à la fin du livre de codes pour trouver une variable appropriée, par exemple LSK_55. L'entrée du livre de codes pour cette variable fournit son concept abrégé, le libellé de la question, l'univers et les fréquences non pondérées. Cela confirme que la variable LSK_55 indique, parmi tous les répondants, ceux qui ont hésité à utiliser la langue minoritaire au cours des cinq dernières années.

La première étape de la production de l'estimation de la population et de son intervalle de confiance (IC) consiste à lier le fichier analytique au fichier des poids bootstrap (voir la section 7 pour plus d'informations sur les poids bootstrap).

```
data SOLMP_ADULTS;

    merge SOLMP.solmp_adt_m SOLMP.solmp_adt_bsw;

    by masterid;

run;
```

Pour calculer le pourcentage requis et son intervalle de confiance (IC), il convient de produire un tableau de fréquences à l'aide de la procédure SURVEYFREQ, comme le montre l'exemple de code SAS suivant.

```
PROC SURVEYFREQ data=SOLMP_ADULTS varmethod=bootstrap;

      WEIGHT weight;

      REPWEIGHTS bsw1-bsw1000 /repcoefs=0.016;

      TABLES LSK_55 / clwt cl (type=wilson adjust=no truncate=yes);

      WHERE LSK_55 IN (1,2);

RUN;
```

Le code ci-dessus spécifie l'option bootstrap comme méthode utilisée pour calculer les estimations de la variance. La variable de poids de l'enquête de l'EPLOSM « weight » est spécifiée dans l'instruction WEIGHT de la procédure, et les poids bootstrap bsw1 à bsw1000 fournis avec les données dans l'instruction REPWEIGHTS. Pour l'EPLOSM, il est extrêmement important d'utiliser le facteur multiplicatif approprié (repcoefs=0,016), souvent appelé « facteur d'ajustement de Fay », pour évaluer l'erreur d'une estimation (voir section 8.6.2).

La ligne TABLES spécifie la variable d'intérêt et l'option « cl » pour obtenir l'intervalle de confiance (IC) des estimations uniquement pour les répondants qui ont répondu oui ou non à la question, comme spécifié par la ligne

WHERE. Comme le pourcentage à estimer est une proportion, le fait de spécifier « type=Wilson » garantit que l'IC fourni par la procédure est l'IC de Wilson modifié et non l'IC de Student, qui est l'option par défaut.

Les résultats arrondis sont présentés dans le tableau ci-dessous.

Tableau B1
Estimations et intervalles de confiance pour les adultes, Canada

	Taille estimée de la population	Intervalles de confiance à 95 % pour la taille de la population		Proportion estimée	Intervalles de confiance à 95 % pour la proportion	
Hésitation à utiliser la langue officielle minoritaire	nombre pondéré arrondi	de	à	pourcentage, basée sur des nombres pondérés non arrondis	de	à
Oui	563 120	539 190	587 040	27,3	26,2	28,5
Non	1 499 100	1 475 160	1 523 039	72,7	71,5	73,8
Total	**2 062 220**	**2 059 680**	**2 064 750**	**100,0**	**…**	**…**

… n'ayant pas lieu de figurer

Note : La procédure SAS SURVEYFREQ calcule les limites de confiance de Wald pour les effectifs pondérés et les limites de confiance de Wilson modifiées pour les pourcentages. Voir la documentation SAS pour plus de détails.

Source: Enquête sur la population de langue officielle en situation minoritaire de 2022.

Selon ce tableau et tel qu'indiqué dans la section 2.2 du rapport de l'EPLOSM, « Au Canada, environ un quart (27 %) des adultes de langue officielle minoritaire ont vécu [...] au cours des cinq années précédentes [...] une situation où ils hésitaient à utiliser la langue officielle de la minorité ». ((Pépin-Filion et al., 2024).

Il est important de noter que pour obtenir le pourcentage dans le tableau, selon les directives d'arrondissement, les chiffres pondérés non arrondis de SAS doivent être utilisés au numérateur et au dénominateur (c.-à-d. 563 116/ 2 062 216 = 27,3 %). L'IC à 95 % autour de cette estimation va d'une limite inférieure de 26,2 % (arrondie) à une limite supérieure de 28,5 % (arrondie).

Veuillez noter que le chiffre non pondéré du numérateur (obtenu à l'aide de PROC SURVEYFREQ et non montré ici) sur lequel cette estimation est basée est bien supérieur à la valeur minimale de 10 requise pour qu'une statistique puisse être diffusée en vertu des règles de confidentialité. Veuillez vous référer à la section 8.1 de ce document pour plus d'informations.

Il est également nécessaire de s'assurer que l'estimation peut être diffusée, avec ou sans avertissement, conformément aux règles de qualité de la diffusion. Le domaine d'intérêt étant le niveau national (tout le Canada), les règles du tableau 8.2 doivent être utilisées. Le nombre non pondéré d'unités au dénominateur (non montré ici) est ≥ 180 et la longueur de l'intervalle de confiance (IC) est de 2,3 points de pourcentage (p.p.), soit moins de 14 p.p. Ainsi, la proportion de 27,3 % et son IC (26,2 % à 28,5 %) peuvent être diffusés sans avertissement (catégorie de diffusion A), et les utilisateurs doivent utiliser l'IC comme indicateur de qualité.

Déterminer si la différence observée entre deux estimations est statistiquement significative

Une fois les IC à 95 % établis, il est relativement simple de déterminer si la différence entre deux estimations est statistiquement significative. Si les deux intervalles ne se chevauchent pas, on peut conclure que les chiffres de population sous-jacente estimés sont différents. En termes plus techniques, l'hypothèse nulle selon laquelle il n'y a pas de différence entre les quantités de population sous-jacente estimées, au niveau de 5 %, peut être rejetée. Toutefois, si les deux intervalles se chevauchent, on ne peut rien conclure. Comme indiqué précédemment dans la section 8.6.5, il s'agit d'une approche prudente et non d'un test exact : le taux de rejet réel est inférieur aux 5 % visés ici. Une façon d'effectuer un test exact serait de créer un IC pour la différence entre les deux proportions, ou d'utiliser la procédure DIFFMEANS de SAS.

Poursuivant l'exemple précédent, supposons qu'un utilisateur souhaite déterminer s'il existe une différence statistiquement significative entre le pourcentage d'adultes de langue anglaise qui ont hésité à utiliser l'anglais au Québec et le pourcentage d'adultes de langue française qui ont hésité à utiliser le français au Canada en dehors du Québec. Le tableau suivant a été créé de la même manière que ci-dessus, en ajoutant une deuxième variable « IN_QUE » et l'option « row » à la ligne TABLES pour produire un tableau croisé comparant l'hésitation à utiliser la langue minoritaire au Québec et à l'extérieur du Québec :

```
PROC SURVEYFREQ data=SOLMP_ADULTS varmethod=bootstrap;

    WEIGHT weight;

    REPWEIGHTS bsw1-bsw1000 /repcoefs=0.016;

    TABLES IN_QUE * LSK_55 / clwt cl row (type=wilson adjust=no truncate=yes);

    WHERE LSK_55 IN (1,2);

RUN;
```

Tableau B2
Estimations et intervalles de confiance pour les adultes, Québec, Canada hors Québec

	Taille estimée de la population	Intervalles de confiance à 95 % pour la taille de la population		Proportion estimée	Intervalles de confiance à 95 % pour la proportion	
Hésitation à utiliser la langue officielle minoritaire	nombre pondéré arrondi	de	à	pourcentage, basée sur des nombres pondérés non arrondis	de	à
Québec						
Oui	324 280	304 880	343 690	30,5	28,7	32,3
Non	720 350	720 350	759 310	69,5	67,7	71,3
Total	**1 064 110**	**1 061 690**	**1 066 530**	**100**	**…**	**…**
Canada hors Québec						
Oui	238 840	225 760	251 920	23,9	22,6	25,3
Non	759 270	746 200	772 340	76,0	74,7	77,4
Total	**998 110**	**997 300**	**998 910**	**100**	**…**	**…**
Total						
Oui	563 120	539 190	587 040	27,3	26,2	28,5
Non	1 499 100	1 475 160	1 523 039	72,7	71,5	73,8
Total	**2 062 220**	**2 059 680**	**2 064 750**	**100**	**…**	**…**

… n'ayant pas lieu de figurer

Note : La procédure SAS SURVEYFREQ calcule les limites de confiance de Wald pour les effectifs pondérés et les limites de confiance de Wilson modifiées pour les pourcentages. Voir la documentation SAS pour plus de détails.

Source : Enquête sur la population de langue officielle en situation minoritaire de 2022.

Selon ce tableau, 30,5 % des adultes de langue anglaise ont hésité à utiliser l'anglais au Québec et l'intervalle de confiance à 95 % correspondant est de 28,7 % à 32,3 % (chiffres arrondis à la première décimale), comparativement à 23,9 % des adultes de langue française qui ont hésité à utiliser le français au Canada à l'extérieur du Québec et l'IC de 22,6 % à 25,3 %.

Pour déterminer si la différence observée entre les deux estimations est statistiquement significative, les deux IC à 95 % peuvent être comparés :

Canada hors Québec : Québec :

22,6 % à 25,3 % 28,7 % à 32,3 %

Comme les deux intervalles ne se chevauchent pas, on peut affirmer, au seuil de signification de 5 %, que la proportion d'adultes de langue anglaise hésitant à utiliser l'anglais au Québec est significativement différente de la proportion d'adultes de langue française hésitant à utiliser le français au Canada à l'extérieur du Québec. Voir la section 8.6.5 pour plus d'explications sur l'utilisation des IC pour déterminer la signification statistique.

Par conséquent, il est possible d'affirmer, comme dans le rapport de l'EPLOSM, que la « proportion était légèrement plus élevée au Québec (31%) qu'au Canada hors Québec (24%) ». (Pépin-Filion et al., 2024).

Annexe C – Une introduction non technique à la méthodologie d'enquête

La présente annexe constitue une introduction générale et non technique à la méthodologie d'enquête permettant de mieux comprendre les éléments méthodologiques propres à l'EPLOSM décrits dans ce document. Un compte rendu plus complet du processus d'enquête est offert par Méthodes et pratiques d'enquête.

Pour l'EPLOSM, une liste de toutes les personnes d'intérêt, appelée base de sondage, a été créée à partir des données recueillies lors du Recensement de la population de 2021, ce qui en fait une enquête postcensitaire. La base a été divisée en sous-groupes mutuellement exclusifs appelés strates, les groupes à l'intérieur desquels est effectué l'échantillonnage et dont sont constitués les domaines d'analyse d'intérêt. Dans chaque strate, un échantillon aléatoire a été tiré ; assemblés, ces échantillons forment l'échantillon (principal) de l'EPLOSM, qui identifie les personnes à contacter lors des activités de collecte de données.

Une fois les données des personnes échantillonnées collectées, vérifiées et anonymisées, elles entrent dans la composition du fichier analytique aux côtés de la variable de poids d'enquête qui encapsule les informations pertinentes sur la conception de l'enquête nécessaires pour formuler des énoncés statistiques (appelées inférences) sur l'ensemble de la population d'intérêt. En essence, l'inférence d'enquête consiste à extrapoler de la partie au tout, car il y a rarement un intérêt analytique pour le sous-groupe de la population constituée des personnes échantillonnées.

Généralement, l'objet d'une inférence est une certaine quantité d'intérêt dans la population (appelée paramètre d'intérêt). La proportion de personnes dans la population qui partagent une caractéristique donnée est un exemple de paramètre d'intérêt. La valeur du paramètre d'intérêt serait connue seulement si des données exactes étaient recueillies auprès de toutes les personnes d'intérêt dans la population. Cependant, en pratique, avec un échantillon, il n'y aura de données que pour un sous-ensemble de ces personnes, et donc la valeur du paramètre dans l'ensemble de la population est inconnue. La question devient alors : comment les données recueillies peuvent-elles être utilisées pour produire une valeur plausible (appelée estimation) pour le paramètre d'intérêt? Et une fois qu'une estimation a été obtenue, dans quelle mesure est-elle plausible ou précise?

Pour extraire une estimation de la masse d'informations contenues dans le fichier analytique, on utilise un outil mathématique appelé estimateur. Un estimateur comporte deux composantes : les données pertinentes à résumer en une seule valeur – l'estimation – et la variable de poids d'enquête, qui assurera que l'estimation basée sur l'échantillon s'aligne avec le paramètre d'intérêt de la population.

L'erreur d'enquête est la différence numérique entre l'estimation obtenue à partir des données et la valeur réelle du paramètre d'intérêt correspondant. Bien que des situations sans erreur d'enquête existent dans les manuels, en pratique il y a toujours des erreurs d'enquête à prendre en compte (les erreurs d'enquête existent même dans un recensement – une enquête où l'échantillon correspond à l'ensemble de la base de sondage – car les réponses fournies peuvent ne pas être exactes, et la base de sondage peut ne pas correspondre à la population inférentielle d'intérêt et donc présenter des problèmes de couverture).

Bien qu'il n'y ait aucun moyen de savoir avec certitude en pratique quelle est l'ampleur de l'erreur d'enquête – c'est-à-dire à quel point l'estimation obtenue à partir des données collectées est éloignée du paramètre d'intérêt – il existe un moyen de savoir si l'estimation est vraisemblablement proche de la valeur réelle, et cela implique d'évaluer le biais et la variance de l'estimateur.

Considérons l'exemple numérique fictif suivant. Supposons que la proportion réelle d'intérêt dans une population soit de 15 % et que l'estimation basée sur les données collectées à partir d'un échantillon aléatoire soit de 12,6 %, ce qui est 2,4 points de pourcentage (pp) en dessous de la valeur réelle. Si un autre échantillon avait été sélectionné, peut-être que l'estimation aurait plutôt été de 13,4 %, ce qui est 1,6 pp en dessous de la valeur réelle. Remarquez également que ces deux estimations, 12,6 % et 13,4 %, sont séparées de 0,8 pp l'une de l'autre.

Avec seulement deux échantillons, on remarque déjà les particularités suivantes quant à l'erreur d'enquête : 1) les estimations obtenues ici ont tendance à être inférieures à la valeur réelle, et 2) elles ne concordent pas mais se situent plutôt à une certaine distance l'une de l'autre. La première remarque suggère un biais dans les estimations - c'est-à-dire la tendance qu'elles ont de se situer du même côté de la valeur réelle - tandis que la seconde témoigne de leur variance, de la tendance de l'estimateur à proposer des valeurs plausibles différentes d'un échantillon à l'autre pour un même paramètre d'intérêt. Pour établir formellement le biais et la variance d'un

estimateur, nous aurions besoin d'examiner tous les échantillons possibles et leurs estimations, pas seulement ces deux-là. Comme en pratique nous ne disposons que des données d'un seul échantillon, cette méthode ne peut être mise en œuvre. L'évaluation du biais et de la variance peut cependant être réalisée en adoptant des pratiques exemplaires d'atténuation du biais et en utilisant une méthode d'estimation de la variance adaptée aux enquêtes.

Naturellement, nous souhaitons que les erreurs d'enquête soient dans leur ensemble aussi faibles que possible étant donné les contraintes opérationnelles. Mais comme il y a deux composantes à prendre en compte – le biais et la variance – la réduction des erreurs d'enquête devient la recherche d'un compromis entre biais et variance. Divers autres exposés relatifs au biais et à la variance dans le contexte d'une enquête existent, y compris les suivants[10].

Les pratiques exemplaires de conduite d'enquêtes et de formulation d'inférences visent à minimiser le biais, ne laissant alors que la variance à évaluer. Il y a de nombreuses raisons de rechercher ce compromis spécifique entre biais et variance. Premièrement, bien que le biais soit un terme statistique, c'est aussi un mot à connotation négative dans le langage courant. Deuxièmement, le biais n'est pas affecté par une augmentation de la taille de l'échantillon : selon toute vraisemblance, l'utilisation de plus de données provenant de la même source ne fera que renforcer l'emprise que le biais a déjà sur les estimations. En revanche, la variance diminue lorsque la taille de l'échantillon augmente, ce qui offre un moyen pratique de réduire l'erreur d'enquête. Enfin, le biais statistique est difficile à évaluer en pratique, car la valeur réelle inconnue d'intérêt est nécessaire pour établir la direction et l'ampleur du biais. Or, la variance peut généralement être évaluée puisqu'elle dépend uniquement de la manière dont les estimations se comportent dans leur ensemble – à quel point elles sont proches ou éloignées les unes des autres – et non de leur position générale par rapport au paramètre (inconnu) d'intérêt.

Dans certains cas, l'erreur d'enquête peut être largement évitée en adoptant les pratiques exemplaires de conduite d'enquête. Par exemple, avant le début de la collecte des données, le questionnaire de l'EPLOSM a fait l'objet d'un examen approfondi et rigoureux pour s'assurer que les questions étaient clairement posées et cohérentes en français et en anglais. Cependant, dans d'autres cas, comme l'échantillonnage et la non-réponse, l'erreur d'enquête se produira malgré les meilleurs efforts de l'équipe d'enquête et les pratiques exemplaires adoptées. L'objectif est alors d'atténuer autant que possible l'erreur d'enquête, ce que l'équipe d'enquête fait en s'appuyant sur des méthodologies éprouvées lors de la production du fichier analytique. Ensuite, en utilisant la variable de poids d'enquête fournie dans l'ensemble de données analytiques, les analystes s'efforcent d'éliminer les biais de leurs inférences, convertissant ainsi ce qui reste de l'erreur d'enquête en variance qui doit être évaluée à l'aide des variables de poids Bootstrap également fournies.

Pour illustrer le compromis biais-variance qui découle de l'utilisation de la variable de poids d'enquête, considérons l'exemple numérique fictif suivant. Supposons que 235 personnes dans une population de mille partagent une caractéristique donnée. Pour estimer ce nombre d'intérêt, un échantillon aléatoire simple sans remise de 500 personnes est tiré, ce qui représente la moitié de la taille de la population. Considérons l'estimateur non pondéré qui retourne simplement comme estimation le nombre de personnes d'intérêt identifiées au sein de l'échantillon, disons 115 pour cet échantillon-ci. Même sans connaître le nombre réel, le biais de l'estimateur non pondéré est évident, puisqu'il ne compte que les personnes d'intérêt dans la moitié de la population : l'autre moitié, celle non sélectionnée, n'est pas prise en compte. Considérons maintenant l'estimateur pondéré qui double le nombre de personnes identifiées dans l'échantillon ; ce facteur de 2 correspond à la valeur du poids d'enquête dans ce cas. Bien qu'elle ne soit pas exempte d'erreur, l'estimation pondérée de 230 est au moins comparable en magnitude à la valeur réelle de 235.

Par conséquent, la variable de poids d'enquête a réduit le biais que l'estimateur non pondéré présentait initialement (en fait, on peut formellement démontrer que l'estimateur pondéré est non biaisé dans ce cas). Cependant, comme une estimation pondérée ici est le double de sa contrepartie non pondérée, les estimations pondérées seront inévitablement plus éloignées les unes des autres que ne le sont les estimations non pondérées. Ainsi, l'élimination du biais que présentent les estimations non pondérées se traduit par un accroissement de la variance au sein des estimations pondérées. L'utilisation de la variable de poids d'enquête constitue un compromis judicieux entre biais

10. Variance et biais et Statistique 101 : biais statistique

et variance, car on force alors toute l'erreur à se présenter sous forme de variance afin qu'elle soit évaluée à l'aide des variables de poids Bootstrap fournies avec les données.

Pourquoi n'y a-t-il qu'une seule variable de poids d'enquête dans le fichier analytique, mais mille variables de poids Bootstrap qui accompagnent les données? Idéalement, la variance exacte serait évaluée en tirant à plusieurs reprises des échantillons de la population, mais cela n'est pas réalisable car une enquête ne peut se permettre de collecter des données que pour un seul échantillon, et la variable de poids d'enquête est attachée à cet unique échantillon. La procédure Bootstrap consiste à sélectionner 1 000 échantillons à partir de l'échantillon principal, ici celui de l'EPLOSM, de telle manière que la variabilité observée dans les 1 000 estimations Bootstrap qui en résultent approximera la variance qui serait observée parmi toutes les estimations possibles.

Une fois que la variance a été évaluée à l'aide des variables de poids Bootstrap fournies, elle doit être rapportée au moyen d'un intervalle de confiance (IC) plutôt que d'un coefficient de variation (CV). En effet, l'expérience et les travaux d'investigation (par exemple, Neusy et Mantel, 2016) ont montré que les CV peuvent traduire de façon trompeuse la précision des estimations de proportions ou de comptes.

Annexe D – Intervalles de confiance des enquêtes postcensitaires de 2022

La présente annexe fournit des renseignements généraux sur les intervalles de confiance, notamment les méthodes utilisées pour les construire pour les enquêtes postcensitaires de 2022, à savoir l'Enquête canadienne sur l'incapacité (ECI), l'Enquête sur la population de langue officielle en situation minoritaire (EPLOSM), l'Enquête auprès des peuples autochtones (EAPA) et le Supplément sur les Inuit du Nunavut (EAPA-SIN). Ces méthodes comprennent celles relatives à l'intervalle de confiance de Wilson modifié pour les statistiques de type proportion et la méthode percentile-bootstrap pour les autres statistiques. Des renseignements sommaires sont également fournis sur les hypothèses sous-jacentes et sur les propriétés de ces méthodes. Les utilisateurs qui souhaitent connaître les détails théoriques sont invités à consulter les documents cités à la fin de cette annexe.

À propos des intervalles de confiance

Les estimations des enquêtes postcensitaires sont accompagnées d'intervalles de confiance pour permettre aux utilisateurs d'effectuer des inférences statistiques valides. Tel que discuté plus en détail dans l'annexe E, l'inférence statistique est le processus qui consiste à tirer des conclusions sur la population à partir des données recueillies auprès des répondants à une enquête. Les estimations d'enquête produites à partir d'un échantillon présentent un certain degré d'incertitude en raison du processus d'échantillonnage — des estimations différentes auraient pu être obtenues si un autre échantillon avait été sélectionné — et de la non-réponse. Cette incertitude doit être prise en compte si l'on veut effectuer une inférence statistique correcte.

Comme on le précise dans la documentation de chaque enquête, des estimations de la variance sont produites pour quantifier l'incertitude des estimations. Ces estimations de la variance peuvent être utilisées pour produire des mesures de la qualité fondées sur la variance, y compris des erreurs types, des coefficients de variation et des intervalles de confiance. Pour les estimations des enquêtes postcensitaires de 2022, on a choisi d'utiliser des intervalles de confiance, car ils indiquent qu'une incertitude entoure les estimations et fournissent une fourchette de valeurs possibles. Ils offrent également une mesure objective de l'erreur d'échantillonnage sous une forme facile à interpréter, car des intervalles de confiance plus larges sont naturellement associés à une plus grande incertitude. De plus, les intervalles de confiance conviennent à tous les types d'estimations, ce qui n'est pas le cas pour certaines autres mesures de la qualité fondées sur la variance.

Un intervalle de confiance est associé à un niveau de confiance, lequel indique dans quelle mesure on peut être certain que le paramètre réel de population se trouve à l'intérieur de l'intervalle de confiance. Le niveau de confiance est souvent formulé comme $1-\alpha$, où α est le niveau de signification du test d'hypothèse correspondant. Pour les enquêtes postcensitaires, le niveau de confiance est fixé à 95 % et α est donc égal à 0,05. Le niveau de confiance correspond à la couverture prévue des intervalles de confiance. En d'autres termes, si le processus qui a généré les données de l'échantillon était répété un très grand nombre de fois et que l'on construisait les intervalles de confiance correspondants pour chaque estimation en utilisant la même méthode, la proportion de ces intervalles de confiance contenant la valeur réelle de la population devrait être proche du niveau de confiance indiqué (ou nominal). En utilisant des études de simulation, les méthodes d'intervalles de confiance peuvent être évaluées pour s'assurer que leur couverture réelle est proche de leur niveau de confiance nominal.

Plusieurs méthodes différentes peuvent être utilisées pour construire des intervalles de confiance. Les méthodes de Wald et de Student sont le plus couramment utilisées et constituent les méthodes par défaut dans la plupart des logiciels statistiques[11]. Ces deux méthodes supposent une distribution connue pour l'estimateur, c'est-à-dire la fonction statistique utilisée pour produire des estimations. La distribution de l'estimateur peut être considérée comme la forme géométrique qui se dessinerait si tous les échantillons possibles étaient sélectionnés de la population et si l'estimation était effectuée pour chacun d'entre eux, produisant une estimation différente à chaque fois. Bien que les intervalles de confiance de Wald et de Student soient faciles à calculer et généralement valides, ils s'appuient sur des hypothèses qui ne tiennent pas dans certains contextes, ce qui peut mener à une sous-couverture.

11. Les intervalles de confiance de Wald et de Student sont décrits dans les sections 7.2 et 7.3 du Rapport technique sur l'échantillonnage et la pondération, Recensement de la population, 2021.

D'autres méthodes, comme l'intervalle de Wilson modifié pour les proportions, font des hypothèses différentes sur la distribution de l'estimateur. Leur couverture améliorée en fait un meilleur choix pour une inférence statistique valide. Fait intéressant, l'intervalle de confiance percentile-bootstrap est obtenu d'une façon complètement différente. Plutôt que de supposer une forme pour la distribution de l'estimateur, la méthode bootstrap utilise la distribution empirique qui provient des estimations calculées à partir de chacune des répliques bootstrap. Les deux méthodes sont décrites plus en détail dans les prochaines sections.

Intervalle de confiance de Wilson modifié

Pour les enquêtes postcensitaires, la méthode d'intervalle de confiance de Wilson modifié est utilisée pour toutes les statistiques de type proportion (p. ex. proportions[12], répartition en pourcentage[13], parts[14]).

La méthode d'intervalle de confiance de Wilson modifié a été choisie en raison de sa couverture généralement supérieure pour les estimateurs de type proportion et des aspects pratiques de sa mise en œuvre. La méthode est basée sur l'intervalle de confiance de Wilson pour un plan d'échantillonnage aléatoire simple avec remise (EASAR) (Wilson, 1927). Pour les enquêtes postcensitaires, on utilise une version modifiée de l'intervalle de confiance de Wilson qui a été adaptée aux plans d'échantillonnage complexes (Kott et Carr, 1997). Des développements théoriques et des études de simulation approfondies ont montré que cette méthode a de bonnes propriétés dans la plupart des situations. Cette méthode donne de meilleurs résultats que les intervalles de confiance de Wald et de Student dans les situations où ces intervalles de confiance présentent une sous-couverture pour les statistiques de type proportion parce que les hypothèses sur lesquelles ils s'appuient ne tiennent pas (Neusy et Mantel, 2016; Statistique Canada, 2023).

Méthode

La borne inférieure et la borne supérieure d'un intervalle de confiance de Wilson modifié de 95 % pour une statistique de type proportion p sont données par :

$$\text{borne inférieure} = \frac{\hat{p} + z^2/2n_e}{1+z^2/n_e} - \frac{z\sqrt{\hat{p}(1-\hat{p}) + z^2/4n_e}}{\sqrt{n_e}\left(1+z^2/n_e\right)}$$

$$\text{borne supérieure} = \frac{\hat{p} + z^2/2n_e}{1+z^2/n_e} + \frac{z\sqrt{\hat{p}(1-\hat{p}) + z^2/4n_e}}{\sqrt{n_e}\left(1+z^2/n_e\right)}$$

où :

- $\hat{p}$ est l'estimation de p;
- z est le $1-\alpha/2 = 97{,}5^{e}$ percentile de la distribution normale standard ;
- $n_e = \min\left(\frac{n}{deff(\hat{p})}, n\right)$ est la taille effective de l'échantillon ;
- $deff(\hat{p}) = \frac{\hat{V}(\hat{p})}{\hat{p}(1-\hat{p})/n}$ est l'effet de plan estimé associé à $\hat{p}$ par rapport à un plan de sondage EASAR ;
- n est la taille de l'échantillon (c.-à-d. le nombre non pondéré de répondants) au dénominateur de la proportion ;
- $\hat{V}(\hat{p})$ est la variance estimée de $\hat{p}$.

12. Une proportion est une relation entre deux quantités, souvent exprimée sous forme de fraction. Elle représente habituellement la comparaison d'une partie avec le tout.
13. La répartition en pourcentage correspond à la ventilation d'un total en ses composantes, exprimée sous forme de pourcentage du total.
14. Une part désigne généralement la partie ou le pourcentage d'une catégorie ou d'un segment particulier dans un ensemble de données. Dans le contexte de la répartition du revenu, une part désigne la portion ou le pourcentage du revenu total détenu par un groupe spécifique.

Quand $\hat{p}=0$ ou $\hat{p}=1$, la valeur pour n_e n'est pas définie. Dans ce cas, les bornes de l'intervalle de confiance de Wilson sont calculées comme suit :

- Pour $\hat{p}=0$, les bornes de l'intervalle de confiance sont $\left(0, \frac{1}{1+n/z^2}\right)$;
- Pour $\hat{p}=1$, les bornes de l'intervalle de confiance sont $\left(\frac{1}{1+n/z^2}, 1\right)$.

Propriétés

Outre le fait qu'il offre une meilleure couverture que les intervalles de confiance de Wald et de Student pour les petites tailles d'échantillons, et lorsque le paramètre de population est près de zéro ou de un, l'intervalle de confiance de Wilson modifié pour une proportion donnée possède la propriété désirable de préservation de l'étendue, ce qui signifie que sa borne inférieure n'est jamais inférieure à zéro et que sa borne supérieure n'est jamais supérieure à un. Étant donné que les proportions ne peuvent pas prendre de valeurs en dehors de l'intervalle de zéro à un, on peut s'attendre à ce que les intervalles de confiance pour les proportions excluent les valeurs négatives et les valeurs supérieures à un.

Il convient également de noter que, contrairement aux intervalles de Wald et de Student, l'intervalle de confiance de Wilson modifié pour les proportions peut être asymétrique, ce qui signifie que l'estimation ne se situera pas exactement au centre de l'intervalle. L'asymétrie est faible lorsque la taille effective de l'échantillon est grande ou lorsque la proportion estimée est près de 0,5.

Tout comme les intervalles de confiance de Wald et de Student, l'intervalle de confiance de Wilson modifié pour les proportions peut faire l'objet d'une certaine sous-couverture, particulièrement lorsque la taille de l'échantillon est très petite, lorsque la valeur de la proportion est près de zéro ou de un, ou lorsqu'il existe une forte corrélation entre les membres d'un même ménage. Néanmoins, la méthode de Wilson modifiée atteint généralement le taux de couverture nominal dans des situations extrêmes, comparativement aux méthodes de Wald et de Student. En général, elle maintient une couverture aussi bonne sinon meilleure que celle des méthodes de Wald et de Student dans ces situations.

Intervalle de confiance percentile-bootstrap

Dans le cas des enquêtes postcensitaires, cette méthode est utilisée pour toutes les statistiques sauf celles de type proportion.

La méthode percentile-bootstrap a été choisie en raison de sa façon directe de dériver des intervalles de confiance pour divers estimateurs complexes de paramètres de population. L'estimation est répétée avec chaque poids de réplique bootstrap et les estimations qui en résultent sont utilisées pour approximer la distribution de l'estimateur au lieu de faire des hypothèses sur sa forme. Les percentiles appropriés de cette distribution approximative sont ensuite utilisés pour délimiter une zone autour de l'estimation qui correspond à un intervalle de confiance de 95 %. Des développements théoriques et des études de simulation ont montré que cette méthode présente de bons résultats dans la plupart des situations et peut être appliquée à des estimateurs complexes (Efron et Tibshirani, 1986; Tibshirani, 1984).

Méthode

La borne inférieure et la borne supérieure d'un intervalle de confiance percentile-bootstrap de 95 % pour une statistique *Y* qui n'est pas de type proportion sont données par :

$$\text{borne inférieure} = \hat{Y} + \sqrt{R}\left(\hat{Y}_{(BI)} - \hat{Y}\right)$$

$$\text{borne supérieure} = \hat{Y} + \sqrt{R}\left(\hat{Y}_{(BS)} - \hat{Y}\right)$$

où :

- $\hat{Y}$ est l'estimation[15];
- R est le facteur d'ajustement bootstrap[16];
- $\hat{Y}_{(BI)}$ est la BI^e estimation bootstrap croissante non manquante;
- $\hat{Y}_{(BS)}$ est la BS^e estimation bootstrap croissante non manquante;
- $BI = \frac{\alpha}{2} \times B$;
- $BS = \left(1 - \frac{\alpha}{2}\right) \times B$;
- α est le niveau de signification;
- B est le nombre de répliques (à l'exclusion du nombre de répliques où $\hat{Y}_j = .$, où $j = 1, ..., B$)[17].

Si BI et BS ne sont pas des nombres entiers, alors $\hat{Y}_{(BI)}$ et $\hat{Y}_{(BS)}$ est la moyenne des deux estimations bootstrap contiguës ascendantes. Pour les enquêtes postcensitaires, les valeurs BI et BS sont respectivement égales à 25 et 975.

Pour les enquêtes postcensitaires, il est essentiel d'appliquer le facteur d'ajustement bootstrap R, sinon la marge d'erreur associée à l'estimation sera sous-estimée. En d'autres termes, la longueur de l'intervalle de confiance sera sous-estimée si le facteur d'ajustement de Fay est omis dans le calcul des bornes inférieure et supérieure.

Propriétés

Contrairement aux intervalles de Wald et de Student, l'intervalle de confiance percentile-bootstrap respecte la transformation, ce qui signifie que la méthode est valide s'il existe une transformation monotone de l'estimateur qui normalise la distribution de l'estimateur. Cette transformation n'a pas besoin d'être connue; elle doit seulement exister. Dans le cas des intervalles de Wald et de Student, une telle transformation devrait être explicitement spécifiée.

De plus, comme l'intervalle de confiance de Wilson modifié, l'intervalle de confiance percentile-bootstrap préserve l'étendue, ce qui signifie que la méthode produit toujours des intervalles qui se situent dans la fourchette de valeurs admissibles pour le paramètre.

Tout comme d'autres intervalles de confiance, l'intervalle de confiance percentile-bootstrap ne donne pas de bons résultats lorsque l'estimateur est biaisé. Un biais dans la distribution bootstrap entraîne un biais dans l'intervalle de confiance. On sait également que la couverture de l'intervalle de confiance percentile-bootstrap tend à être inférieure au taux nominal lorsque la taille de l'échantillon est petite. Néanmoins, il parvient généralement à un meilleur équilibre des côtés gauche et droit par rapport aux intervalles de Wald et de Student (Efron et Tibshirani, 1994).

15. $\hat{Y}$ peut être un total, un quantile, la différence de deux estimations, etc.
16. *R* est souvent appelé le facteur d'ajustement de Fay pour les enquêtes postcensitaires. Pour l'ECI, l'EAPA et l'EPLOSM, R = 16.
17. Il y a B=1000 répliques bootstrap pour chaque enquête postcensitaire.

Annexe E – Introduction à l'analyse de données d'enquêtes complexes

E.1 Introduction

L'analyse de données d'enquêtes complexes[18] peut sembler une tâche intimidante, avec de nombreux pièges qui guettent l'analyste non averti. En effet, les techniques statistiques classiques et les logiciels que les utilisateurs connaissent généralement ne sont pas tous adaptés à l'utilisation de données d'enquêtes complexes et doivent souvent être modifiés. Par exemple, malgré son nom, le poids d'enquête a peu à voir avec la variable de poids utilisée dans une régression linéaire *pondérée*, une technique destinée aux données non issues d'enquêtes que l'on trouve couramment dans les logiciels statistiques.

Les analystes doivent utiliser les variables de poids d'enquête et de poids Bootstrap fournies ainsi qu'un logiciel avec des procédures adaptées aux enquêtes pour effectuer leurs analyses sur les données de l'EPLOSM. Dans le cas de SAS, cela signifie recourir à sa suite de procédures SURVEY plutôt qu'à ses procédures classiques; par exemple, il faut utiliser PROC SURVEYREG au lieu de PROC REG pour ajuster un modèle de régression linéaire aux données de l'EPLOSM.

Heureusement, de nombreuses options existent pour l'analyste qui souhaite se familiariser davantage avec l'analyse des données d'enquête, notamment des ouvrages tels que Lohr (2021), Heeringa et al. (2020), Lumley (2010) et Lewis (2016), ainsi que du matériel et des tutoriels en ligne comme ceux-ci[19]. De plus, il est fortement recommandé aux analystes de solliciter des conseils et des avis auprès d'un analyste expérimenté ou d'un statisticien spécialisé dans les enquêtes tout au long du processus analytique.

C'est en effet un processus analytique, car l'analyse des données d'enquête ne se limite pas à déterminer quelles méthodes et quels logiciels utiliser ; par exemple, Heeringa et al. (2020) identifient et discutent les six étapes suivantes qu'un plan analytique devrait couvrir pour produire une analyse statistiquement valide des données d'enquête qui soit significative à la fois pour les communautés sociales et de recherche :

1. Définir le problème et énoncer les objectifs
2. Comprendre le plan d'échantillonnage
3. Comprendre les variables du plan, les construits sous-jacents et les données manquantes
4. Analyser les données
5. Interpréter et évaluer les résultats de l'analyse
6. Rapporter les estimations et les inférences tirées des données d'enquête

Que le plan analytique soit réalisé uniquement à des fins exploratoires ou pour soutenir la prise de décision basée sur les données, les variables de poids d'enquête et de poids Bootstrap doivent être utilisées avec les données de l'EPLOSM pour traiter l'erreur d'enquête.

Cette annexe présente une introduction brève aux tests statistiques, en commençant par une discussion sur le rôle que doivent jouer les variables de poids d'enquête et de poids Bootstrap dans toute diffusion d'informations statistiques issues des données de l'EPLOSM.

E2. Gérer l'erreur d'enquête par le biais des variables de poids d'enquête et de poids Bootstrap

L'erreur d'enquête renvoie à la différence entre une quantité inconnue d'intérêt, par exemple, la proportion de personnes dans la population ayant une certaine caractéristique, et son estimation basée sur les données d'enquête. Étant donné que les données d'enquête ne sont normalement collectées que pour un sous-ensemble de toutes les personnes concernées (par exemple, les répondants d'un échantillon de la population) et que l'expérience d'enquête d'un répondant peut ne pas fournir les réponses exactes aux questions posées, il s'avère qu'en pratique toute estimation est issue d'informations incomplètes et imparfaites. Par conséquent, malgré les

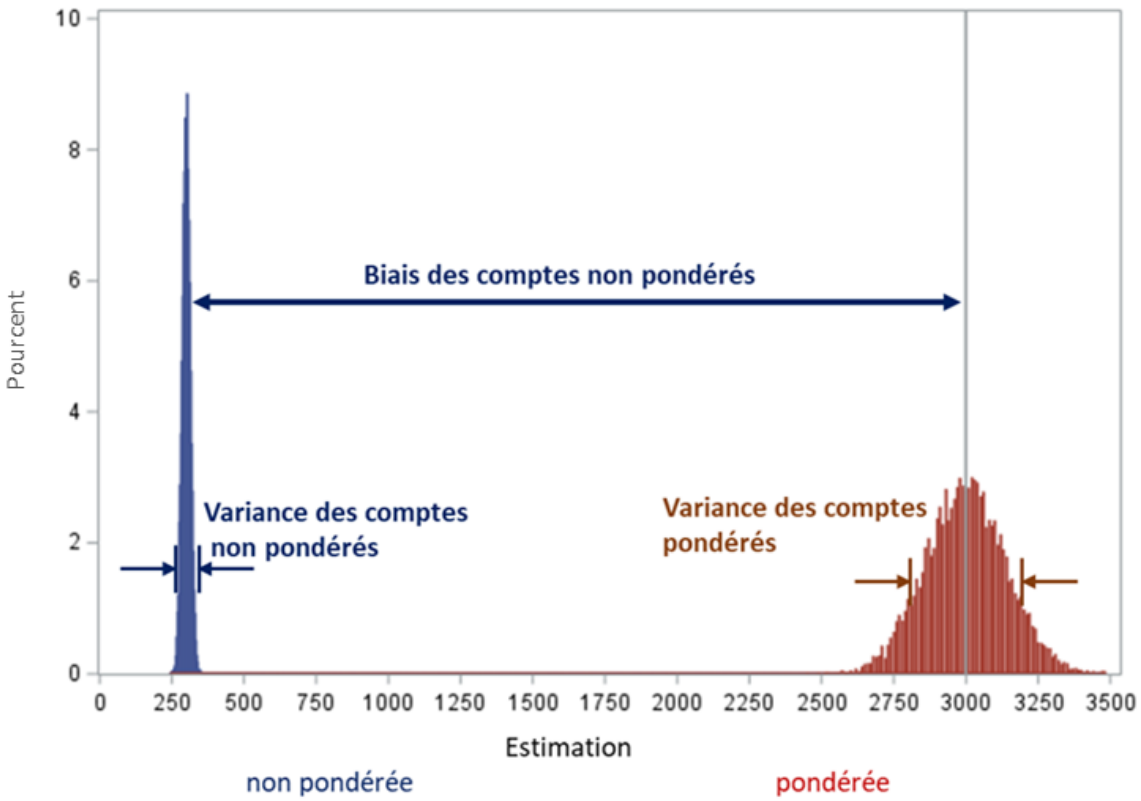
18. Les données d'enquête « complexes » sont des données collectées par l'entremise d'un plan d'échantillonnage comportant une combinaison ou une autre des éléments suivants : stratification, mise en grappe, poids d'enquête de valeurs inégales ou facteurs de correction pour la population finie.
19. Catalogue d'apprentissage de la formation de la littératie des données

meilleurs efforts de l'équipe d'enquête et les pratiques exemplaires adoptées, il y a toujours des erreurs d'enquête à prendre en compte lors de la diffusion d'informations statistiques.

L'erreur d'enquête admet deux composantes, le biais (statistique) et la variance, qui sont expliquées en termes non techniques ici[20]. La section 3.0 et l'annexe C soulignent déjà que la variable de poids d'enquête doit être utilisée lors de la production d'une estimation pour réduire son biais au minimum, ne laissant que sa variance – l'autre composante de l'erreur d'enquête – à évaluer en utilisant les variables de poids Bootstrap fournies avec les données de l'EPLOSM.

Le graphique suivant décrit les distributions d'estimations pondérées et non pondérées du nombre d'individus d'intérêt tirées de l'échantillonnage répété d'une population simulée; il permet de comparer leurs biais et variances.

Figure E1.1
Distributions des comptes pondérés et non pondérés



Source : Image de l'auteur.

Le biais des comptes non pondérés est évident : le centre de la distribution bleue se trouve loin du véritable compte d'intérêt indiqué la ligne verticale grise. Ce biais était prévisible puisqu'un compte non pondéré n'est rien d'autre que le dénombrement des individus d'intérêt au sein de l'échantillon, sans tentative de compenser pour l'absence des individus d'intérêt non sélectionnés. Par contraste, les comptes pondérés (qui s'obtiennent en sommant le poids d'enquête des individus d'intérêt de l'échantillon) sont sans biais : le centre de la distribution rouge coïncide avec le vrai compte, la ligne verticale grise.

On constate dans ce graphique que le recours au poids d'enquête a converti[21] le biais que présentent les estimations non pondérées en variance supplémentaire pour les estimations pondérées non biaisées : la distribution en rouge est plus large que celle en bleu. Parce que le biais est difficile (voire impossible) à évaluer en pratique, sa

20. Variance et biais (statcan.gc.ca)
21. Un énoncé plus précis serait que la variable de poids d'enquête a éliminé une partie de l'erreur d'échantillonnage initialement présente dans les estimations non pondérées et a converti l'erreur résiduelle en variance au sein des estimations pondérées.

magnitude dépendant du compte d'intérêt (inconnu) qu'on cherche à estimer, il est avantageux que toute l'erreur des estimations pondérées se présente sous forme de variance, car celle-ci s'évalue à l'aide des poids Bootstrap fournies avec les données.

Certains analystes s'opposent à l'utilisation indiscriminée de la variable de poids d'enquête afin d'obtenir des estimations non biaisées à partir de données d'enquêtes complexes. Ils affirment, à juste titre, qu'il existe des situations où l'utilisation du poids d'enquête ne protège pas contre le biais et ne fait alors qu'augmenter la variance associée à l'estimation. Par conséquent, ils soutiennent que dans de tels cas, une estimation non pondérée d'une proportion serait plus appropriée qu'une estimation pondérée.

Le raisonnement qu'ils adoptent est le suivant : lorsque la variable d'intérêt est indépendante du plan d'échantillonnage et des méthodes de collecte utilisées, l'estimation pondérée d'une proportion n'aura pas moins de biais que son homologue non pondéré, mais une variance possiblement plus importante. Bien que cela soit correct, il est généralement très difficile (voire impossible) en pratique de tirer profit de telles opportunités lorsqu'elles se présentent. D'une part, on ne s'attend pas à ce que les analystes aient une connaissance approfondie de la méthodologie d'enquête en général, ni de la façon dont elle a été spécifiquement appliquée à l'EPLOSM, pour défendre de manière concluante l'hypothèse d'indépendance de l'analyse par rapport au processus d'enquête. De plus, selon le contexte analytique, les informations d'enquête nécessaires pour étayer une telle affirmation pourraient ne pas être disponibles dans le fichier analytique. Pour ces raisons, il est recommandé aux analystes d'utiliser la variable de poids d'enquête dans toute situation analytique afin de préserver leurs résultats du biais induit par le processus d'enquête, même si dans certaines situations une estimation plus précise pourrait être obtenue sans utiliser le poids de l'enquête.

E3. Deux éléments clés à un test statistique

Un test statistique est une technique spécialisée qui s'utilise dans le cadre d'un plan analytique pour étudier la cohérence d'une hypothèse de travail avec les données collectées. Par exemple, l'hypothèse pourrait être émise que la proportion de personnes dans la population ayant une certaine caractéristique est la même pour deux groupes d'intérêt.

Il y a deux éléments clés à un test statistique : une statistique et sa loi de référence associée. Le fichier analytique d'une enquête constitue un ensemble vaste et complexe d'informations. Le but de la statistique est de résumer les informations du fichier jugées pertinentes dans le contexte analytique en une seule valeur numérique. Dans l'exemple précédent, un test statistique naturel à mener implique la statistique t calculée à partir des proportions estimées pour chacun des groupes et de leurs variances.

Mais en raison de l'erreur d'enquête, même lorsque les proportions des deux groupes sont égales dans la population, on ne peut pas s'attendre à ce que la valeur de la statistique calculée à partir des données de l'enquête soit nécessairement zéro, bien qu'elle devrait être proche de zéro.

En présence d'erreur d'enquête, il est naturel de se demander : dans quelle mesure la valeur de la statistique doit-elle s'écarter de zéro pour rejeter avec confiance l'hypothèse d'égalité des proportions au sein de la population? C'est là qu'intervient la loi de référence associée à la statistique, le deuxième élément clé d'un test statistique. La loi indique quelles valeurs sont plausibles pour la statistique de test lorsque l'hypothèse est vraie en tenant compte de l'erreur d'enquête.

E4. Niveau de signification : tracer une ligne dans le continuum de valeurs plausibles

Comme on peut s'y attendre, toute valeur admissible pour la statistique d'un test est plausible (ou compatible) à un certain degré avec l'hypothèse à l'étude. Néanmoins, dans ce continuum de plausibilité, il faut bien tracer une ligne quelque part, identifiant un seuil au-delà duquel la valeur d'une statistique est considérée comme trop improbable pour que l'hypothèse de travail soit cohérente avec les données.

Le seuil est normalement fixé de manière que, si l'hypothèse était effectivement vraie – c'est-à-dire que, peu importe à quel point la valeur statistique calculée à partir des données est extrême, on saurait avec certitude qu'elle est cohérente avec l'hypothèse – alors l'analyste accepterait de rejeter à tort l'hypothèse 5 % du temps. Autrement dit, l'analyste saurait que le test mène au rejet de l'hypothèse (pourtant vraie) dans 5 % des échantillons réalisables. (Évidemment, l'analyste ne saura pas en pratique si l'unique échantillon tiré par l'enquête fait partie de ce 5% des échantillons.) Cette probabilité de rejeter à tort une hypothèse vraie est appelée le niveau de signification du test, et le rôle de la loi de référence est de transformer cette probabilité en une valeur seuil pour la statistique du test.

De façon complémentaire, malgré l'incertitude engendrée par l'erreur d'enquête, l'analyste éviterait de rejeter[22] une hypothèse vraie 95 % du temps ; cette probabilité est appelée le niveau de confiance du test. Cette notion rappelle les intervalles de confiance, et ce n'est pas une coïncidence puisque les tests statistiques et l'estimation d'un paramètre de population sont tous deux ancrés dans le même cadre statistique. En fait, nous verrons ci-dessous comment un test statistique peut être réalisé en utilisant un intervalle de confiance.

Un test statistique est valide lorsque son niveau de signification effectif correspond au niveau visé (ou nominal). Ainsi, un test qui s'avère rejeter une hypothèse vraie donnée 40 % du temps n'est pas valide au niveau visé habituel de 5 %, mais il serait valide si 40 % était réellement le niveau de signification visé. Dans ce dernier cas, cependant, un résultat statistiquement significatif n'aurait que très peu de valeur scientifique étant donné un taux de faux positifs aussi élevé : des fluctuations normales dans la composition des données en raison de l'aléa de la sélection suffiraient à déclencher un rejet par le test de l'hypothèse lorsque celle-ci est pourtant vraie.

Il est à noter que bien que 5 % soit un choix courant, un niveau de signification plus bas de 1 % peut être nécessaire lorsque l'analyste est confronté à une charge de preuve analytique plus importante, par exemple pour éclairer certaines décisions politiques.

E5. Détermination du caractère significatif d'un résultat statistique

Un plan analytique bien défini peut impliquer le calcul de la valeur d'une statistique à partir des données, qui sera comparée à la valeur-seuil fournie par la loi de référence appropriée et correspondant au niveau de signification visé, disons 5%. Le résultat associé est considéré comme statistiquement significatif au niveau visé lorsque la valeur calculée pour le test est plus extrême que la valeur-seuil.

Cependant, en pratique, il arrive que certains se fient à la valeur-*p* fournie de manière routinière par les logiciels statistiques avec la valeur calculée de la statistique afin de déterminer son caractère significatif. La valeur-*p* est la probabilité d'obtenir une valeur au moins aussi extrême que celle qui vient d'être calculée pour la statistique lorsque l'hypothèse sous-jacente est vraie. Ainsi, la valeur-*p* correspond au niveau de signification lorsque la statistique calculée est égale à la valeur-seuil.

En somme, l'utilisation d'une valeur-*p* implique la comparaison de deux probabilités – la valeur-*p* et le niveau de signification visé – plutôt que de comparer deux valeurs de la statistique comme auparavant (celle calculée à partir des données avec la valeur-seuil correspondant au niveau visé).

Bien que les valeurs-*p* aient toujours été populaires, elles peuvent être mal utilisées et mal interprétées. Par exemple, si une faible valeur-*p* révèle effectivement quelque chose sur l›hypothèse de travail, elle ne dit rien sur la force d›une hypothèse concurrente considérant les mêmes données. En effet, peut-être que la seule autre hypothèse ayant du sens est encore moins cohérente avec les données que l›hypothèse de travail rejetée.

22. Bien que cela revienne à accepter l'hypothèse, dans la pratique statistique on n'accepte jamais pleinement une hypothèse comme étant la vérité; on dit plutôt, comme c'est le cas ici, qu'on ne trouve pas d'indications dans les données de la rejeter.

Aussi pratique que cela puisse paraître, se contenter de chercher des résultats avec de petites valeurs-*p* à rapporter ne remplace pas le travail de fond nécessaire pour concevoir et exécuter un plan d'analyse bien défini et pertinent. En fait, la pratique délibérée de produire en masse des résultats statistiques dans le seul but de rapporter comme significatifs ceux ayant de faibles valeurs-p constitue un problème d'inférence connu sous le nom de p-hacking[23].

E6. Caractérisation de la loi de référence

Puisque la validité d'une conclusion statistique dépend de la loi de référence, il est important d'assigner la bonne loi à la statistique pour que le niveau de signification effectif corresponde à celui visé. En pratique, il n'y a souvent pas qu'une seule loi de référence associée à une statistique donnée, mais plutôt toute une famille de lois.

Par exemple, le test t mentionné précédemment est associé à la famille de lois statistiques connue sous le nom de lois de Student : ces lois sont analytiquement identiques à l'exception d'un paramètre qu'elles ont en commun, dont la valeur détermine l'étalement de chacune. Ce paramètre est communément appelé le nombre de degrés de liberté associés à la loi. Dans le cas de la loi de Student, plus le nombre de degrés de liberté est grand, plus la forme de la loi est resserrée et donc plus l'intervalle de valeurs qu'elle juge plausible pour la statistique est étroit. Par conséquent, une loi plus étroite conduira à un rejet plus fréquent – à juste titre ou non – qu'une loi plus large. Par conséquent, spécifier la bonne valeur du paramètre – c'est-à-dire identifier la taille correcte de l'intervalle des valeurs plausibles sous le niveau de signification visé – est primordial pour obtenir un test statistique valide.

Cependant, spécifier le nombre correct de degrés de liberté à utiliser avec des données d'enquête complexes peut s'avérer une tâche difficile, car divers aspects de la conception du plan de sondage, autres que la seule taille de l'échantillon, jouent un rôle. En pratique, on s'appuie couramment sur la règle approximative qui, dans le cas de l'EPLOSM, fait correspondre le nombre de degrés de liberté à la taille d'échantillon disponible pour la sous-population d'intérêt, moins le nombre de strates impliquées dans l'analyse. (Cependant, comme il n'y a aucun moyen d'identifier les strates à partir du fichier analytique de l'EPLOSM, il est recommandé d'utiliser « 1 » comme nombre de strates impliquées dans l'analyse.)

Avertissement : de nombreux logiciels, par exemple SAS, font correspondre le nombre de degrés de liberté au nombre de poids Bootstrap fournis avec les données. Étant donné qu'il y a 1 000 poids Bootstrap pour l'EPLOSM, cela surestimerait le nombre correct de degrés de liberté à utiliser dans toutes les analyses, sauf celles impliquant de grands domaines. Ainsi, les analyses pour les domaines de petite à moyenne taille risquent d'obtenir des résultats appuyant de manière trop forte le rejet d'hypothèses. Les analystes doivent examiner la documentation du logiciel pour l'option de spécifier leur propre nombre de degrés de liberté.

E7. Conduite d'un test statistique à l'aide d'un intervalle de confiance

Tel que mentionné déjà, les tests statistiques sont étroitement liés aux intervalles de confiance. Par exemple, on peut rejeter au niveau de signification de 5 % l'hypothèse de proportions égales entre deux groupes lorsque l'intervalle de confiance de (100-5) % = 95 % associé à la différence estimée des proportions ne contient pas 0. En somme, l'intervalle de confiance correspond à la zone de non-rejet du test associé.

Remarque : cette pratique n'est pas équivalente à celle de rejeter lorsque les deux intervalles de confiance 95 % associés aux proportions estimées ne se chevauchent pas; cette dernière approche est plus conservatrice que le test usuel puisqu'elle rejette moins souvent, incluant dans les cas où la différence entre les proportions est véritablement nulle.

Au lieu de recourir à un intervalle de confiance Student, il est possible d'obtenir un intervalle de confiance par la méthode des percentiles Bootstrap – voir l'annexe D pour les détails. Une caractéristique particulière de la méthode Bootstrap est qu'elle utilise sa propre loi de référence, établie à partir des données. Puisque la loi employée par le Bootstrap n'est pas déterminée par l'analyste (que ce soit directement ou indirectement en raison des paramètres par défaut du logiciel), il y a moins de chances qu'elle soit incorrectement spécifiée.

23. Data dredging — Wikipédia (wikipedia.org)

E8. Dernières remarques

Ce court exposé avait pour but de sensibiliser à l'emploi des variables de poids d'enquête et de poids Bootstrap au sein de leurs analyses, de même qu'à certains enjeux susceptibles de survenir lors de la conduite de tests statistiques à partir de données d'enquête. L'intention était de fournir les bases formatives nécessaires à l'exploration autonome de la littérature sur le sujet et à la consultation auprès d'analystes d'expérience ou des statisticiens d'enquête, afin d'établir avec succès un plan analytique probant. Parmi les points à retenir, on relève notamment :

- La conception d'un plan analytique en consultation avec un analyste d'expérience ou un statisticien d'enquête;
- L'emploi des variables de poids d'enquête et de poids Bootstrap afin de composer adéquatement avec l'erreur d'enquête;
- L'obtention d'un test statistique valide exige que le niveau de signification réalisé corresponde à celui visé et rapporté – Ne pas identifier correctement la loi de référence et effectuer des comparaisons multiples sont des situations propices à la réalisation d'un test non valide;
- Le recours sans discernement aux valeurs-*p* facilement mises à la disposition de l'analyste par les logiciels statistiques afin de tirer des conclusions à partir des données ne constitue pas une pratique recommandée.

Bibilographie

Efron, B. et Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Statistical Science, Vol. 1, No. 1, 54-77.

Efron, B. et Tibshirani, R. J. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC.

Gagné, C., G. Roberts et L.-A. Keown. « Estimation pondérée et estimation de la variance bootstrap pour analyser des données d'enquête : Comment les effectuer dans certains logiciels choisis? » Le Bulletin technique et d'information des centres de données de recherche. Hiver 2014. Vol. 6 No.1 Statistique Canada No 12-002-X au catalogue — No 2014001

Heeringa, S.G., B.T. West et P.A. Berglund. 2020. Applied Survey Data Analysis. Second Edition. CRC Press.

Korn, E.L., et Graubard, B.I. (1998). « Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête ». *Techniques d'enquête*, 24, 209-218.

Kott, P.S. et Carr, D.A. (1997). "Developing an Estimation Strategy for a Pesticide Data Program." Journal of Official Statistics, Vol. 13, No. 4, 367-383.

Lewis, T.H. 2016. Complex Survey Data Analysis with SAS. CRC Press.

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

Lohr, S. 2021. Sampling: Design and Analysis. Third Edition. CRC Press.

Lumley, T. 2010. Complex Surveys – A Guide to Analysis Using R. Wiley.

Neusy, E. et H. Mantel. (2016). "Confidence Intervals for Proportions Estimated from Complex Survey Data". Proceedings of the Survey Methods Section of the Statistical Society of Canada Annual Meeting, June 2016.

Pépin-Filion, D., L. Cornelissen et É. Lemyre, 2024, « Situation des populations de langue anglaise au Québec et de langue française au Canada hors Québec – Résultats de l'Enquête sur la population de langue officielle en situation minoritaire de 2022 », Série thématique sur l'ethnicité, la langue et l'immigration, produit no 89-657-X 2024008 au catalogue de Statistique Canada.

Savard, Sarah-Anne (2023). Overview of the quality component of the dissemination strategy for the 2022 Canadian Survey on Disability, October 2023, Document interne.

Statistique Canada. 2020. GTAB specifications V21. Document interne. Ottawa, Ontario.

Statistique Canada. 2023. Enquête canadienne sur l'incapacité, 2022 : Guide des concepts et méthodes Catalogue no. 89-654-X2023004. Ottawa, Ontario.

Statistique Canada. 2023. Evaluating confidence interval methods for the 2021 Census using the Census Monte Carlo simulation environment. Document interne. Ottawa, Canada.

Statistique Canada. 2023. Rapport technique sur l'échantillonnage et la pondération, Recensement de la population. Catalogue no. 98-306-X2021001. Ottawa, Ontario. (accessed December 12, 2023).

Statistique Canada. 2024. Enquête canadienne sur l'incapacité de 2022 : Guide de l'utilisateur des fichiers de données analytiques. Document interne disponible dans les Centres de données de recherche. Ottawa, Ontario.

Tibshirani, Robert (1984). Bootstrap confidence intervals, Department of Statistics Stanford University, Technical report, No. 3.

Wilson, E.B. 1927. "Probable Inference, the Law of Succession, and Statistical Inference." Journal of the American Statistical Association, Vol. 22, no. 158. p. 209-212.