Guide de l'utilisateur des microdonnées

Enquête sur la population active (EPA)

Fichier de microdonnées à grande diffusion (FMGD)

Janvier 2025

Statistics Canada Statistique Canada

Canada

Table des matières

1.0 Introduction 3
2.0 Concepts et Définitions 4
3.0 Méthodologie de l'enquête 5
3.1 Population cible 5
3.2 Échantillonnage 5
3.3 Collecte de données 5
3.4 Vérification et modification 6
3.5 Création de variables dérivées 6
3.6 Pondération 6
3.7 Révisions 7
4.0 Contrôle de la divulgation 7
5.0 Tabulation, analyse et diffusion 8
5.1 Estimation et analyse statistique 8
5.2 Interprétation de la variance 9
6.0 Poids bootstrap de Poisson pour l'estimation de la variance 10
6.1 Comment créer les poids bootstrap de Poisson 10
6.2 Comment utiliser les poids bootstrap pour calculer la variance 11
6.2.1 Comment utiliser les poids bootstrap pour calculer les CV 12
6.2.2 Comment utilizer les poids bootstrap pour créer les intervalles de confiance 12
Annexe A – Formules 14
A1 : Création des poids bootstrap Poisson 14
A2 : Indicateurs de qualité 14
Annexe B – Domaines de calage recommandés 16
Annexe C – Examples 17
C1 : Générer les poids bootstrap 17
C2 : Estimer la variance en utilisant les poids bootstrap. 19
Annexe D – Exemples de programmes 21
D1 : Exemple de programme du logiciel SAS 21
D2 : Exemple de programme du logiciel R 25
Références 29

1.0 Introduction

L'Enquête sur la population active (EPA) est une enquête mensuelle menée auprès des ménages par Statistique Canada. Depuis sa création en 1945, l'EPA a pour objectif de répartir la population en âge de travailler en trois catégories relatives au marché du travail qui s'excluent mutuellement, à savoir celles des personnes occupées, des chômeurs et des inactifs, ainsi que de fournir des données descriptives et explicatives sur chacun de ces groupes. Les données de l'enquête fournissent des renseignements sur les grandes tendances du marché du travail, par exemple sur les mouvements observés relativement à l'emploi entre les différents secteurs industriels, sur les heures travaillées, sur le taux d'activité et le taux de chômage du marché du travail.

Ce fichier de microdonnées à grande diffusion (FMGD) renferme des données non agrégées associées à une vaste gamme de variables recueillies dans le cadre de l'Enquête sur la population active (EPA). Ce produit est destiné aux utilisateurs qui préfèrent réaliser leur propre analyse et qui s'intéressent à des sous-groupes particuliers de la population ou qui désirent effectuer des recoupements de variables ne faisant pas partie de nos produits catalogués. Les données ont été modifiées pour veiller à ce qu'aucune personne ou entreprise ne puisse être identifiée directement ou indirectement. Les variables les plus susceptibles de conduire à l'identification d'un individu sont supprimées du fichier de microdonnées ou sont regroupées dans des catégories plus larges.

Ce guide a été élaboré pour faciliter l'utilisation du FMGD et l'interprétation des résultats. Pour des informations plus détaillées sur l'EPA et sa méthodologie, veuillez-vous référé au *Guide de l'Enquête sur la population active* (71-543-G).

Toute question concernant les fichiers de microdonnées ou leur utilisation doit être adressée à :

Statistique Canada
Centre de l'information sur le marché du travail
Courriel: **statcan.labour-travail.statcan@statcan.gc.ca**

2.0 Concepts et Définitions

Les concepts et définitions d'emploi et de chômage adoptés par l'Enquête sur la population active reposent sur ceux qui sont approuvés par l'Organisation internationale du travail (OIT).

Les concepts d'emploi et de chômage trouvent leurs origines dans la théorie de l'offre de travail comme facteur dans la production. Dans ce contexte, la production correspond aux biens et services figurant dans le Système de comptabilité nationale. C'est pourquoi les travaux ménagers sans rémunération et le travail bénévole ne sont pas considérés comme du « travail » aux fins de l'EPA, bien que ces activités ne diffèrent pas nécessairement du travail rémunéré, que ce soit sur le plan de l'objectif ou de la nature des tâches à accomplir.

Les **personnes occupées** comprennent :

1. Les personnes qui, durant la semaine de référence de l'EPA, ont effectué au moins une heure de travail rémunéré, pour le compte d'un employeur ou à leur propre compte[1] ; **ET**
2. Les personnes qui, durant la semaine de référence de l'EPA, avaient un emploi ou une entreprise, mais n'étaient pas au travail.[2]

Les **chômeurs** comprennent :

1. Les personnes qui, durant la semaine de référence, n'avaient pas d'emploi ou d'entreprise, étaient disponibles pour travailler et avaient cherché un emploi au cours des quatre dernières semaines se terminant par la semaine de référence ; **ET**
2. Les personnes qui, durant la semaine de référence, n'avaient pas d'emploi ou d'entreprise, étaient disponibles pour travailler et avaient été mises à pied temporairement à cause de la conjoncture économique, mais s'attendaient à être rappelées au travail ; **ET**
3. Les personnes qui, durant la semaine de référence, n'avaient pas d'emploi ou d'entreprise, étaient disponibles pour travailler et devaient commencer un nouvel emploi au cours des quatre semaines suivant la période de référence :
   (Note : Les personnes dans les catégories 2 et 3 peuvent ne pas avoir cherché un emploi au cours des quatre semaines se terminant par la semaine de référence.)

Les **personnes inactives** comprennent:

1. Les personnes qui n'étaient ni occupées, ni au chômage durant la semaine de référence et qui voulaient travailler, mais qui ne satisfaisaient pas aux critères relatifs aux chômeurs décrits ci-dessus ; **ET**
2. Les personnes qui étaient incapables de travailler ou n'étaient pas disponibles pour travailler ou celles qui ne désiraient pas travailler.

Dans le FMGD de l'EPA, la variable "situation dans la population active" (LFSSTAT) indique si un individu est employé, au chômage ou ne fait pas partie de la population active.

[1] Les personnes occupées comprennent aussi les personnes qui ont fait un travail familial non rémunéré, qui est défini comme un travail non rémunéré qui contribue directement à l'exploitation d'une ferme, d'une entreprise ou d'un cabinet de professionnels appartenant à un membre apparenté du même ménage et exploité par lui.

[2] Cette catégorie n'inclut pas les personnes mises à pied temporairement ou entre deux emplois occasionnels, ainsi que celles qui devaient commencer un emploi à une date ultérieure.

Une liste complète des variables du FMGD de l'EPA se trouve dans le dictionnaire de données du FMGD de l'EPA.

3.0 Méthodologie de l'enquête

Cette section fournit un bref aperçu de la méthodologie de l'EPA afin d'aider les utilisateurs du FMGD. Pour des informations plus détaillées sur la méthodologie de l'EPA, veuillez vous référer au *Guide de l'Enquête sur la population active* (71-543-G).

3.1 Population cible

La population cible de l'Enquête sur la population active (EPA) comprend toutes les personnes de 15 ans et plus dont le lieu de résidence habituel se situe au Canada, ce qui inclut les résidents non permanents (RNP) — c'est-à-dire les titulaires d'un permis de travail ou d'études, les membres de leur famille, les demandeurs d'asile, les personnes protégées et les groupes apparentés — ainsi que les résidents permanents (immigrants reçus) et les personnes nées au Canada.

Les populations exclues de la population cible de l'EPA comprennent les personnes vivant dans les réserves, les membres à temps plein des Forces armées régulières et les pensionnaires d'établissements (notamment les personnes détenues dans les pénitenciers et les patients d'hôpitaux et d'établissements de soins infirmiers). Les personnes exclues de l'enquête représentent moins de 2 % de la population de 15 ans et plus.

3.2 Échantillonnage

L'échantillon de l'EPA est tiré d'une base de sondage de zone et repose sur un plan stratifié à plusieurs degrés utilisant un échantillonnage probabiliste. L'échantillon mensuel de l'EPA compte environ 68 000 ménages, ce qui se traduit par la collecte de données sur le marché du travail concernant environ 100 000 personnes.

L'EPA utilise un plan d'échantillonnage par panel tournant. Dans les provinces, les logements sélectionnés restent dans l'échantillon de l'EPA pendant six mois consécutifs. Chaque mois, environ un sixième des logements de l'échantillon de l'EPA en est à son premier mois d'enquête, un sixième en est à son deuxième mois d'enquête, et ainsi de suite. Ces six échantillons indépendants sont appelés groupes de rotation. Sur le FMGD de l'EPA, les identifiants des répondants (REC_NUM) sont attribués de manière aléatoire chaque mois ; il n'est donc pas possible de calculer les flux du marché du travail ou de suivre un répondant tout au long de ses six mois d'appartenance à l'échantillon.

3.3 Collecte de données

Les questionnaires de l'EPA portent sur les activités de chaque membre du ménage pendant la semaine de référence de l'EPA, qui est généralement la semaine contenant le 15e jour du mois. Les entrevues de l'EPA sont réalisées chaque mois au cours des dix jours qui suivent immédiatement la semaine de référence.

Pendant la période de collecte des données, des intervieweurs formés par Statistique Canada communiquent avec des ménages sélectionnés, soit par téléphone, soit en personne. Les entrevues peuvent également être réalisées en ligne, les membres des ménages sélectionnés recevant une invitation à remplir leur questionnaire sur une plateforme sécurisée de collecte de données de Statistique Canada. Ces entrevues sont réalisées sans l'intervention d'un enquêteur.

Dans chaque logement, les renseignements sur tous les membres du ménage sont habituellement obtenus auprès d'un membre du ménage bien informé. Cette déclaration par procuration, qui représente environ 65 % de l'information recueillie, est utilisée pour éviter les coûts élevés et les délais prolongés qu'entraîneraient les appels répétés nécessaires pour obtenir l'information directement de chaque répondant.

Le questionnaire de l'EPA est disponible en ligne à Questionnaire de l'Enquête sur la population active .

3.4 Vérification et modification

À la fin de la collecte, une série de vérifications est effectuée afin d'identifier et d'éliminer les enregistrements potentiellement en double et d'exclure les enregistrements de non-réponse et hors du champ de l'enquête. Cette étape de vérification permet également d'identifier et de corriger les réponses incohérentes ou non valides.

L'imputation est le processus qui consiste à remplacer les données invalides ou manquantes par des valeurs valides. Les nouvelles valeurs sont fournies de manière à préserver la structure sous-jacente des données et à garantir que les enregistrements obtenus passeront toutes les vérifications requises. En d'autres termes, l'objectif n'est pas de reproduire les vraies valeurs des microdonnées, mais plutôt d'établir des enregistrements de données cohérents à l'interne qui produisent de bonnes estimations agrégées.

Les méthodes d'imputation utilisées dans le cadre de l'EPA comprennent l'imputation par report, l'imputation déterministe et l'imputation par la méthode du plus proche voisin. Dans certains cas, la non-réponse complète – lorsque toutes les données du questionnaire pour un ménage sont manquantes – est résolue par un ajustement pour la non-réponse, comme décrit dans la sous-section intitulée Pondération.

3.5 Création de variables dérivées

La plupart des variables du FMGD de l'EPA sont dérivées en combinant des réponses au questionnaire ou en effectuant des calculs fondés sur ces réponses. Par exemple, le questionnaire de l'EPA demande aux répondants de déclarer leur salaire horaire ou le salaire fixe, y compris les pourboires et les commissions, avant impôt et autres déductions. Le salaire horaire (HRLYEARN) est ensuite calculé en fonction du nombre habituel d'heures de travail rémunérées par semaine.

Les utilisateurs de données doivent consulter le dictionnaire des données du FMGD de l'EPA pour connaître les codes et l'univers de chacune des variables dérivées.

3.6 Pondération

Les données de l'EPA sont pondérées pour qu'il soit possible de totaliser les estimations aux niveaux d'agrégation national, provincial et infraprovincial.

Le plan d'échantillonnage détermine un certain nombre de facteurs de pondération devant servir au calcul des poids individuels. La principale composante est l'inverse de la probabilité de sélection, aussi appelée « poids de base ». Par exemple, dans un secteur où 2 % des ménages sont inclus dans l'échantillon, chaque ménage se voit attribuer un poids de base de 50 (soit 1 ÷ 0,02). Le poids de base est ensuite ajusté pour tenir compte de tout sous-échantillonnage susceptible d'être observé du fait de

l'expansion que pourrait avoir connue le secteur, ainsi que de la non-réponse et des erreurs de couverture.

Dans le cadre de l'EPA, une partie de la non-réponse de l'enquête est compensée au moyen de l'imputation : méthodes du report, de remplacement ou d'imputation par donneur. La non-réponse résiduelle est prise en compte par correction des poids attribués aux ménages répondants du même secteur. Cette manière de procéder repose sur l'hypothèse selon laquelle les caractéristiques des ménages répondants ne diffèrent pas de façon marquée de celles des ménages non-répondants.

Les poids font l'objet d'une correction finale qui sert à tenir compte des erreurs de couverture et à réduire la variabilité échantillonnale des estimations. Les sous-poids sont corrigés au moyen d'un calage composite pour accroître l'efficacité des estimations en tirant profit du chevauchement entre deux échantillons mensuels consécutifs, et pour s'assurer que les estimations de l'enquête sont conformes aux totaux de contrôle par âge, genre et géographie.

3.7 Révisions

Les données de l'EPA sont également ajustées tous les cinq ans à la suite de la diffusion des nouvelles estimations démographiques basées sur le dernier recensement. Depuis janvier 2025, les estimations de l'EPA ont été ajustées pour tenir compte des chiffres de population du recensement de 2021, les révisions remontant à 2011.

Parfois, les données de l'EPA sont révisées pour intégrer des mises à jour des classifications d'industrie et de profession, ou des améliorations méthodologiques, de traitement des données et des systèmes technologiques.

Toutes révisions de l'EPA sont décrites dans l'article « Améliorations apportées à l'Enquête sur la population active (EPA) » (71F0031X)

4.0 Contrôle de la divulgation

La loi interdit à Statistique Canada de rendre publique toute donnée susceptible de révéler de l'information obtenue en vertu de la *Loi sur la statistique* et se rapportant à toute personne, entreprise ou organisation reconnaissable sans que cette personne, entreprise ou organisation le sache ou y consente par écrit. 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.

Ainsi, les données contenues dans les fichiers de microdonnées à grande diffusion (FMGD) peuvent différer des fichiers maîtres de l'enquête détenus par Statistique Canada. Ces différences résultent généralement des mesures prises pour protéger l'anonymat des personnes interrogées. Les mesures les plus courantes sont la suppression d'éléments de données et le regroupement de valeurs dans des catégories plus larges

Par exemple, le fichier maître de l'EPA contient les indicateurs géographiques pour les régions intraprovinciales détaillés, qui inclus les régions métropolitaines de recensement (RMR), agglomérations de recensement, régions économiques et subdivisions de recensement. Les variables géographiques sur

le FMGD de l'EPA sont limitées aux provinces et les neuf RMR le plus grands. De plus, les variables d'âge, industrie et profession ont été groupés sur le FMGD pour protéger la confidentialité des répondants. De plus, certains enregistrements du FMGD ont été perturbés pour renforcer la sécurité. Parfois, un enregistrement individuel du FMGD peut ne pas correspondre à un enregistrement vrai sur le fichier maître de l'EPA.

Des mesures ont été prises pour s'assurer que les estimations issues du FMGD restent proches des estimations officielles calculées à partir du fichier de données de base ; toutefois, il peut y avoir des écarts par rapport aux estimations publiées sur le site Web de Statistique Canada, en particulier pour les petits domaines. En cas d'écart, les estimations figurant dans les tableaux publiés et les autres produits de données sur le site Web de Statistique Canada doivent être considérées comme des statistiques officielles.

5.0 Tabulation, analyse et diffusion

5.1 Estimation et analyse statistique

L'EPA est basée sur un plan d'échantillonnage complexe, comprenant la stratification, plusieurs étapes de sélection et des probabilités inégales de sélection des répondants. L'utilisation de données provenant d'enquêtes aussi complexes pose des défis pour les analystes, car le plan d'enquête et les probabilités de sélection influent sur les procédures d'estimation et de calcul de la variance qui doivent être utilisées. Pour garantir des résultats adéquats, les poids d'enquête fournis (FINALWT) doivent être utilisés pour toutes les estimations, tabulations et analyses statistiques utilisant le FMGD de l'EPA.

Pour les petits domaines, il peut être nécessaire de combiner des fichiers mensuels pour produire des estimations fiables. Pour l'EPA, il est généralement recommandé d'utiliser une moyenne mobile sur trois mois ou une estimation de la moyenne annuelle pour les petits domaines tels que les RMR ou les immigrants. Pour calculer une moyenne mobile sur trois mois, les utilisateurs doivent regrouper les trois FMGD mensuels dans un fichier et diviser le poids de l'enquête par 3. Par exemple, une estimation de la moyenne mobile sur trois mois pour avril serait basée sur les FMGD de février, mars et avril et le poids utilisé serait WT_3MMA = FINALWT/3. De même, pour calculer une estimation annuelle, les utilisateurs doivent regrouper les douze FMGD mensuels et diviser le poids de l'enquête par 12, c'est-à-dire WT_ANNUAL = FINALWT/12. Le fichier combiné et le poids ajusté peuvent ensuite être utilisés pour la tabulation ou l'analyse.

Avant de procéder à toute analyse de l'ensemble de données, il est essentiel de consulter le dictionnaire des données du FMGD de l'EPA. Ce dernier décrit le jeu de codes et l'univers pour chaque variable du fichier de données, ainsi que des notes sur le formatage spécial, les inclusions ou les exclusions.

Par exemple, les variables du FMGD décrivant les heures (UHRSMAIN, AHRSMAIN, UTOTHRS, ATOTHRS, HRSAWAY, PAIDOT, UNPAIDOT, XTRAHRS) et les salaires (HRLYEARN) sont présentées sous forme de nombres entiers avec des décimales implicites. Par conséquent, les données de ces variables doivent être transformées pour garantir une utilisation correcte. Par exemple, le total des heures travaillées par semaine dans tous les emplois comporte une décimale implicite et doit être divisé par 10 pour obtenir une valeur en heures, c'est-à-dire qu'une valeur de 435 pour ATOTHRS correspond à 43,5 heures par semaine. Le salaire horaire moyen comporte deux décimales implicites et doit être divisé par 100 pour

obtenir une valeur en dollars et en centimes, c'est-à-dire qu'une valeur de 2345 pour HRLYEARN correspond à 23,45 $.

Les utilisateurs de données doivent également examiner l'univers de chaque variable avant de calculer des estimations ou d'effectuer des analyses. Par exemple, l'univers de la permanence de l'emploi (PERMTEMP) est « Occupés, employés ». Par conséquent, avant d'effectuer une analyse à l'aide de cette variable, la population d'intérêt doit être limitée aux employés, c'est-à-dire (LFSSTAT = 1 ou LFSSTAT = 2) et (COWMAIN = 1 ou COWMAIN = 2).

5.2 Interprétation de la variance

Il est important de déterminer la qualité de toute estimation utilisée dans une analyse statistique. Bien que la qualité des données soit affectée à la fois par les erreurs d'échantillonnage et les erreurs non liées à l'échantillonnage, les lignes directrices sur la qualité de ce guide de l'utilisateur traitent des évaluations de la qualité déterminées uniquement en fonction de l'erreur d'échantillonnage. Pour de plus amples renseignements sur les erreurs d'échantillonnage et les erreurs non liées à l'échantillonnage, reportez-vous à la Méthodologie de l'Enquête sur la population active du Canada (71-526-x).

Deux facteurs principaux doivent être pris en compte pour déterminer la qualité d'une estimation : le nombre de répondants qui contribuent au calcul de l'estimation et la variabilité échantillonnale de l'estimation.

Pour produire une estimation de qualité acceptable, au moins 5 répondants doivent contribuer au calcul de l'estimation.

Pour la plupart des estimations pondérées, une mesure appropriée de la qualité peut être déterminée en calculant le coefficient de variation (CV) de l'estimation en divisant l'erreur type de l'estimation par l'estimation elle-même, puis en suivant les lignes directrices du tableau ci-dessous.

Tableau 5 : Lignes directrices sur la qualité

Qualité de l'estimation	Lignes directrices
1) Acceptable	Les estimations ont une taille d'échantillon de 5 ou plus, et les coefficients de variation sont moins de 15 %. Aucune mise en garde est nécessaire.
2) Marginale	Les estimations ont une taille d'échantillon de 5 ou plus, et les coefficients de variation sont entre 15 % et 35 %. Les estimations doivent être accompagnées d'un avertissement aux utilisateurs concernant leur qualité.
3) Inacceptable	La taille d'échantillon est moins de cinq, ou les coefficients de variation excèdent 35 %. Statistique Canada recommande de ne pas publier des estimations de qualité inacceptable.

Les estimations de qualité marginale ou inacceptable doivent être accompagnées d'un avertissement destiné à mettre en garde les utilisateurs ultérieurs.

Il convient de noter que pour les estimations de petits ratios, tels que le taux de chômage, le CV peut être faussement élevé. Pour les petits ratios, les différences ou tout autre type d'estimation, un

intervalle de confiance peut fournir une meilleure représentation de la qualité des données. Pour plus d'informations sur le calcul d'un intervalle de confiance à l'aide du FMGD, voir la section 6.2.2..

6.0 Poids bootstrap de Poisson pour l'estimation de la variance

Le calcul d'estimations précises de la variance nécessite une connaissance détaillée du plan d'échantillonnage de l'enquête. Pour protéger la confidentialité des répondants, ce niveau de détail ne peut être inclus dans un FMGD. Pour les utilisateurs des FMGD de l'EPA, la méthode du bootstrap de Poisson peut être utilisée pour estimer la valeur réelle de la variance (Beaumont & Patak, 2012). Si nécessaire, des estimations plus précises de la variance pour la plupart des statistiques qui tiennent compte des complexités du plan d'échantillonnage de l'EPA peuvent être demandées sur la base du recouvrement des coûts en utilisant les poids Bootstrap de Rao-Wu et le fichier maître de l'enquête.

Le bootstrap de Poisson est un cas particulier du bootstrap généralisé. Son caractère pratique pour les utilisateurs, sa flexibilité et sa robustesse font du bootstrap une technique largement acceptée pour l'estimation de la variance par les utilisateurs de l'enquête PUMF. Compte tenu du grand nombre d'observations sur le FMGD de l'EPA et d'assez de rééchantillonnages bootstrap, les hypothèses d'utilisation de la technique bootstrap de Poisson sont facilement satisfaites. Cette technique est un plan unidirectionnel et suppose que toutes les observations sont indépendantes les unes des autres. La méthode proposée utilise l'une des implémentations les plus simples du bootstrap de Poisson, comme décrit dans Beaumont et Patak (2012).

6.1 Comment créer les poids bootstrap de Poisson

Pour chaque unité *k* sur le FMGD, calculer un facteur d'ajustement comme suit :

$$facteur\ d'ajustement_k = 1 + facteur\ poisson_k\ * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}} \qquad (1)$$

où $facteur\ poisson = 1\ ou - 1\ avec\ une\ probabilité\ de\ 50\ \%$ et $finalwt_k$ est le poids de l'enquête pour unité *k* sur le FMGD.

Ensuite, calculer le poids bootstrap de la manière suivante:

$$poids\ bootstrap = finalwt *\ facteur\ d'ajustement \qquad (2)$$

Les équations (1) et (2) peuvent être combinées pour obtenir la formule suivante:

$$poids\ bootstrap = finalwt + \left(facteur\ poisson * finalwt * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}} \right)$$

Cette équation peut également être écrite sous la forme :

$$poids\ bootstrap = finalwt\ \pm finalwt * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}}$$

où le « + » ou « - » est déterminé par le « poisson factor » aléatoire.

Répéter cette procédure pour créer 1 000 rééchantillonnages bootstrap.

Une version calibrée du bootstrap de Poisson peut conduire à des estimations de variance plus proches de celles estimées par le fichier principal de l'EPA. La calage peut être effectué en ajustant chacun des 1 000 poids bootstrap non calibrés à l'aide de :

$$poids\ bootstrap\ calibré = \frac{somme\ de\ finalwt\ par\ domain}{somme\ de\ poids\ bootstrap\ par\ domain} * poids\ bootstrap \quad (3)$$

Où les $poids\ bootstrap$ sont les 1 000 poids de (2), et les sommes de poids bootstrap sont calculés pour chaque rééchantillonnage bootstrap.

Essentiellement, chaque rééchantillonnage bootstrap est calibrée pour correspondre aux estimations faites à l'aide de la variable FINALWT dans le PUMF. La façon la plus efficace de procéder est de choisir des domaines qui sont similaires à ceux utilisés dans l'étalonnage du fichier principal : la province, le groupe d'âge et le sexe. Voir l'annexe B pour les domaines de calage recommandés et l'annexe C.1 pour un exemple montrant comment utiliser les formules pour générer des poids bootstrap.

Ces poids peuvent ensuite être utilisés pour calculer les variances de la même manière que celle décrite dans le bootstrap de Rao-Wu-Yue, qui se trouve au chapitre 7 de la *Méthodologie de l'Enquête sur la population active du Canada* (71-526-x).

6.2 Comment utiliser les poids bootstrap pour calculer la variance

Un fichier des poids bootstrap de Poisson peut être bootstrap weights peut être fusionné avec le FMGD correspondant et utilisé pour calculer la variance de n'importe quelle estimation en suivant les étapes ci-dessous :

1. Calculer l'estimation de l'enquête en utilisant la variable d'intérêt et le poids final de l'enquête (FINALWT).
2. Calculer une estimation bootstrap pour chaque rééchantillonnage bootstrap en utilisant la variable d'intérêt et les poids bootstrap calibrés pour chaque rééchantillonnage. Il devrait en résulter 1 000 estimations bootstrap.
3. Calculer la variance bootstrap des 1 000 estimations bootstrap à l'aide de la formule :

$$bootstrap\ variance(estimation) = \frac{(bs\ estimation_1 - estimation)^2 + \cdots + (bs\ estimation_{1000} - estimation)^2}{1000}$$

où $bs\ estimate_{1-1000}$ sont les estimations calculées à l'aide des 1 000 rééchantillonnages bootstrap.

Si des estimations sont nécessaires par domaine, tel que la province ou le groupe d'âge, les étapes ci-dessus seront mises en œuvre séparément pour chaque niveau du domaine souhaité.

Cette variance peut être utilisée pour calculer les CV ou les intervalles de confiance. Voir l'annexe C2 pour un exemple montrant comment utiliser les poids bootstrap pour produire des indicateurs de qualité.

6.2.1 Comment utiliser les poids bootstrap pour calculer les CV

Une fois la variance calculée comme indiqué ci-dessus, le coefficient de variation peut être calculé simplement comme suit :

$$CV = \frac{\sqrt{bootstrap\ variance(estimation)}}{estimation}$$

Voir la section 5.2 pour plus d'informations sur l'interprétation du CV.

6.2.2 Comment utilizer les poids bootstrap pour créer les intervalles de confiance

Bien que les coefficients de variation soient largement utilisés, l'intervalle de confiance d'une estimation constitue une mesure plus intuitive de l'erreur d'échantillonnage. Un intervalle de confiance constitue une déclaration sur le niveau de confiance que la valeur réelle de la population se situe dans une fourchette de valeurs spécifiée. Par exemple, un intervalle de confiance de 95 % peut être décrit comme suit :

> Si l'échantillonnage de la population est répété indéfiniment, chaque échantillon conduisant à un nouvel intervalle de confiance pour une estimation, alors dans 95 % des échantillons l'intervalle couvrira la vraie valeur de la population.

En utilisant l'erreur-type d'une estimation, les intervalles de confiance peuvent être obtenus en supposant qu'en cas d'échantillonnage répété de la population, les diverses estimations obtenues pour une caractéristique de la population sont normalement distribuées autour de la valeur réelle de la population. Dans cette hypothèse, il y a environ 68 % de chances que la différence entre l'estimation d'un échantillon et la valeur réelle de la population soit inférieure à une erreur-type, environ 95 % de chances que la différence soit inférieure à deux erreurs-type et environ 99 % de chances que la différence soit inférieure à trois erreurs-type. Ces différents degrés de confiance sont appelés niveaux de confiance.

Les intervalles de confiance d'une estimation peuvent être calculés directement en utilisant la formule suivante pour les convertir en intervalles de confiance :

$$(estimation - t * \sqrt{variance(estimation)},\ \ estimation + t * \sqrt{variance(estimation)})$$

où t est le point final approximatif de la distribution normale. En fonction du niveau de confiance, le tableau suivant peut être utilisé :

Tableau 6 : Valeurs critiques pour déterminer les intervalles de confiance

Valeur critique (t)	**Niveau de confiance**
1.0	Intervalle de confiance à 68%
1.6	Intervalle de confiance à 90%
2.0	Intervalle de confiance à 95%
2.6	Intervalle de confiance à 99%

Une autre méthode pour calculer les intervalles de confiance à l'aide des poids bootstrap consiste à déterminer les percentiles de la distribution des estimations correspondant au niveau de confiance

souhaité. Pour ce faire, il faut d'abord trier les estimations des répliques bootstrap par ordre croissant. Ensuite, pour obtenir l'intervalle de confiance, il faut calculer les deux percentiles nécessaires pour délimiter uniformément le niveau de confiance souhaité, comme le montre l'équation suivante :

$$borne\ inférieur\ percentile = \frac{100\ \% - niveau\ de\ confiance}{2}$$

$$borne\ supérieur\ percentile = 100\ \% - \left(\frac{100\ \% - niveau\ de\ confiance}{2}\right)$$

Par exemple, pour créer un intervalle de confiance de 95 %, calculez le 2,5e percentile ((100 % - 95 %) /2) et le 97,5e percentile (100 % - ((100 %-95 %) /2)), déterminant ainsi les limites pour le centre de 95 % des estimations. De même, pour un intervalle de confiance de 68 %, il faut calculer le 16e centile et le 84e centile. Pour trouver les valeurs correspondant à ces percentiles, multipliez le percentile inférieur puis le percentile supérieur par le nombre d'estimations bootstrap et prenez l'estimation correspondante dans les estimations triées. Par exemple, pour un intervalle de confiance de 95 % utilisant 1 000 répliques bootstrap, prenez la 25e estimation et la 975e estimation lorsque les estimations sont triées par ordre croissant.

Note : Les directives de validation qui s'appliquent à l'estimation s'appliquent également à l'intervalle de confiance. Par exemple, si l'estimation n'est pas diffusable, l'intervalle de confiance ne l'est pas non plus.

Voir l'annexe C2 pour un exemple montrant comment utiliser les poids bootstrap pour produire des indicateurs de qualité tels que le CV et les intervalles de confiance.

Annexe A – Formules

A1 : Création des poids bootstrap Poisson

Désignez le nombre de poids de rééchantillonnage bootstrap, *B*. Pour chaque *b* de 1 à *B*, on définit

$$a_{kb} = 1 + \widetilde{a_{kb}}\sqrt{\frac{(w_k - 1)}{w_k}} \qquad (1)$$

Où $\widetilde{a_{kb}} = \begin{cases} 1 \text{ avec probabilité de } 0.5 \\ -1 \text{ avec probabilité de } 0.5 \end{cases}$

w_k est le poids d'enquête de l'unité *k* sur le FMGD.

Ensuite, calculez chaque poids bootstrap w_{kb}^* comme

$$w_{kb}^* = a_{kb} w_k \qquad (2)$$

Pour calibrer les poids bootstrap, ajustez les poids en utilisant

$$bw_{kb} = \frac{n_d}{m_{db}} w_{kb}^* \qquad (3)$$

Où w_{kb}^* est le poids bootstrap *b* de l'unité *k* à partir de (2)

n_d est la somme des poids d'enquête w_k dans le domaine *d*

m_{db} est la somme des poids bootstraps w_{kb}^* dans le domaine *d*

A2 : Indicateurs de qualité

Variance :

$$\widehat{var}(\hat{X}) = \frac{\sum_{b=1}^{1000}(\hat{X}^{*(b)} - \hat{X})^2}{1000}$$

Où $\hat{X}^{*(b)}$ est l'estimation bootstrap du rééchantillonnage bootstrap *b*

$\hat{X}$ est l'estimation de l'enquête

Erreur-type :

$$\hat{\sigma} = \sqrt{\widehat{var}(\hat{X})}$$

Coefficient de variation :

$$CV = \frac{\hat{\sigma}}{\hat{X}}$$

Intervalle de confiance :

$$CI = \hat{X} \pm t * \hat{\sigma}$$

Où t est une approximation de la limite de la distribution normale pour le niveau de confiance requis

Valuer critique (t)	**Niveau de confiance**
1.0	intervalle de confiance à 68 %
1.6	intervalle de confiance à 90 %
2.0	intervalle de confiance à 95 %
2.6	intervalle de confiance à 99 %

OR

$$\left(\frac{100\,\% - p}{2}\right)^{ieme} percentile, \quad \left(100\,\% - \frac{100\,\% - p}{2}\right)^{ieme} percentile$$

Où *p* est le niveau de confiance requis

percentile est calculé en fonction des 1 000 estimations bootstrap

Annexe B – Domaines de calage recommandés

Les domaines optimaux pour calibrer les poids bootstrap sont définis par le croisement de ces trois variables, donnant un total de 220 domaines de calibration :

Province		Groupe d'âge		Genre	
Variable FMGD	Déscription	Variable FMGD	Déscription	Variable FMGD	Déscription
prov = 10	Terre-Neuve-et-Labrador	age_6 = 1	15 à 16 ans	Gender = 1	Hommes+
prov = 11	Île-du-Prince-Édouard	age_6 = 2	17 à 19 ans	Gender = 2	Femmes+
prov = 12	Nouvelle-Écosse	age_12 = 2	20 à 24 ans		
prov = 13	Nouveau-Brunswick	age_12 = 3	25 à 29 ans		
prov = 24	Québec	age_12 = 4	30 à 34 ans		
prov = 35	Ontario	age_12 = 5 or age_12 = 6	35 à 44 an		
prov = 46	Manitoba	age_12 = 7 or age_12 = 8	45 à 54 ans		
prov = 47	Saskatchewan	age_12 = 9	55 à 59 ans		
prov = 48	Alberta	age_12 = 10	60 à 64 ans		
prov = 59	Colombi-Britannique	age_12 = 11	65 à 69 ans		
		age_12 = 12	70 ans et plus		

Annexe C – Examples

C1 : Générer les poids bootstrap

Cet exemple simple utilisera un fichier qui contient 4 unités d'échantillon.

1. Pour chaque unité, créez 1 000 répliques avec un facteur de Poisson égal à 1 ou -1, assigné de manière aléatoire avec une probabilité de 50 %.

Rec_num	FINALWT	Poisson facteur$_1$	Poisson facteur$_2$	Poisson facteur$_3$	...	Poisson facteur$_{1000}$
1	500	1	-1	-1		1
2	450	-1	-1	1		-1
3	150	1	-1	1		1
4	250	-1	1	-1		-1

2. Appliquer la formule (1) : $facteur\ ajustement_k = 1 + facteur\ poisson\ _k * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}}$

Rec_num	FINALWT	Facteur aj.$_1$	facteur aj.$_2$	facteur aj.$_3$	...	facteur aj.$_{1000}$
1	500	= 1+ $1 * \sqrt{\frac{500-1}{500}}$ = 1.99900	= 1+ $-1 * \sqrt{\frac{500-1}{500}}$ = 0.00100	0.00100		1.99900
2	450	= 1+ $-1 * \sqrt{\frac{450-1}{450}}$ = 0.00111	0.00111	1.99889		0.00111
3	150	1.99666	0.00333	1.99666		1.99666
4	250	0.00200	1.99800	0.00200		0.00200

3. Appliquer formule (2): $bootstrap\ weight = finalwt * adjustment\ factor$

Rec_num	FINALWT	Bootstrap poids$_1$	Bootstrap poids$_2$	Bootstrap poids$_3$	...	Bootstrap poids$_{1000}$
1	500	= 500*1.9990 = 999.500	=500*0.0010 = 0.50025	0.50025		999.500
2	450	= 450*0.0011 = 0.50027	0.50027	899.500		0.50027
3	150	299.499	0.50083	299.499		299.499
4	250	0.50050	499.499	0.50050		0.50050

4. Pour l'étape suivante, des données sur la province/âge/genre ont été ajoutées à l'exemple de jeu de données. Calculez les sommes de FINALWT et de tous les poids bootstrap groupés par combinaisons de province/âge/genre.

Rec_num	Prov	Age	Genre	FINALWT	Bootstrap $poids_1$	Bootstrap $poids_2$	Bootstrap $poids_3$	...	Bootstrap $poids_{1000}$
1	**10**	**1**	**1**	500	999.500	0.50025	0.50025		999.500
2	**10**	**1**	**1**	450	0.50027	0.50027	899.500		0.50027
3	**59**	**1**	**1**	150	299.499	0.50083	299.499		299.499
4	**59**	**1**	**1**	250	0.50050	499.499	0.50050		0.50050

Prov	Age	Genre	Somme FINALWT	Somme bootstrap $poids_1$	Somme bootstrap $poids_2$	Somme bootstrap $poids_3$	...	Somme bootstrap $poids_{1000}$
10	1	1	= 500+450 = 950	=999.500+0.50027 = 1,000.0027	1.00052	900.0003		1,000.00027
59	1	1	= 150+250 = 400	=299.499+0.50050 = 299.9995	499.9998	299.9995		299.9995

Appliquez les ratios aux poids bootstrap pour les domaines correspondants afin de créer les poids bootstrap calibrés finaux en utilisant la formule (3) :

$$Poids\ bootstrap\ calibrés = \frac{somme\ finalwt\ par\ domaine}{somme\ poids\ bootstrap\ par\ domaine} * poids\ bootstrap$$

Rec_num	Prov	Age	Genre	bw_1	bw_2	bw_3	...	bw_{1000}
1	10	1	1	=(950/1,000.0027)*999.5 = **949.5247**	**474.99054**	**0.528042**		**949.5247**
2	10	1	1	**0.475256**	**475.0095**	**949.472**		**0.475256**
3	59	1	1	= (400/299.9995)*299.499 **= 399.3327**	**0.400664**	**399.3327**		**399.3327**
4	59	1	1	**0.667334**	**399.5993**	**0.667334**		**0.667334**

C2 : Estimer la variance en utilisant les poids bootstrap.

Cet exemple simple permettra d'estimer la variance du nombre total de chômeurs (LFSStat = 3) en utilisant le jeu de données suivant :

Rec_num	LFSStat	FINALWT	bw_1	bw_2	bw_3	…	bw_{1000}
1	1	500	704.5136	0.646157	0.596452		1547.675
2	1	450	0.352623	0.646182	1072.481		0.774643
3	3	400	563.5404	1032.688	0.596524		0.774705
4	4	200	281.5933	516.0197	476.3259		0.7752
5	1	150	399.3327	0.400664	399.3327		399.3327
6	3	250	0.667334	399.5993	0.667334		0.667334

Notez que dans les calculs de l'exemple, les résultats seront calculés uniquement à partir des 4 répliques bootstrap montrées. En pratique, toutes les 1 000 répliques bootstrap doivent être utilisées.

1. Calculer l'estimation de l'enquête en utilisant les poids d'enquête :

$$\hat{X} = 1 * 400 + 1 * 250 = 650\ chômeurs$$

2. Calculer les estimation bootsrap en utilisant les poids boostrap :

$$\hat{X}^{*(1)} = 563.5404 + 0.667334 = 564.2078\ chômeurs$$

$$\hat{X}^{*(2)} = 1032.688 + 399.5993 = 1432.28725\ chômeurs$$

$$\hat{X}^{*(3)} = 0.596524 + 0.667334 = 1.263858\ chômeurs$$

…

$$\hat{X}^{*(1000)} = 0.774705 + 0.667334 = 1.442039\ chômeurs$$

3. Calculer la variance

$$\widehat{var}(\hat{X}) = [(564.2078 - 650)^2 + (1432.28725 - 650)^2 + (1.263858 - 650)^2 + \cdots + (1.442039 - 650)^2] / 1000$$

$$= 1{,}460.819654$$

Écart-type:

$$\widehat{std}(\hat{X}) = \sqrt{1{,}460.819654}$$

$$= 38.22067 \ldots$$

Coefficient of variation:

$$CV = 38.22067 / 650$$

$$= 0.0588$$

$$= 5.9\%\ (qualité\ acceptable)$$

Intervalle de confiance à 95%:

$$\mathrm{CI} = 650 \pm 2 * 38.22067$$

$$= 650 \pm 74.9125 \ldots$$

$$= (573.56, 726.44)$$

En utilisant les lignes directrices du Tableau 5, le CV pour cette estimation (5,9 %) se situerait dans la plage de qualité acceptable ; toutefois, la qualité réelle de l'estimation serait inacceptable car elle est basée sur une taille d'échantillon de 2 enregistrements.

Annexe D – Exemples de programmes

D1 : Exemple de programme du logiciel SAS

```
/*=========================================================================*/
/* CODE EXEMPLE POUR LE BOOTSTRAP POISSON SUR LE LFS PUMF                  */
/* Ce programme crée un fichier de 1 000 poids bootstrap poisson à utiliser  */
/* avec le LFS PUMF, et montre également quelques exemples de la manière de  */
/* calculer la variance de certaines estimations                           */
/*-------------------------------------------------------------------------*/
/* Input_data = un fichier SAS LFS PUMF                                    */
/* seed = numéro afin de pouvoir reproduire les résultats aléatoires       */
/*=========================================================================*/
%let seed = 123;
%let b = 10; /* Nombre de répliques bootstrap. Utilisez 10 à des fins de test et
d'apprentissage. */
 /* Utilisez 1 000 pour la production des estimations de variance ; */

/*---------CRÉER LES POIDS BOOTSTRAP-------------------------------*/

/* Pour préparer la calibration des poids bootstrap, définissez les
   groupes d'âge de calibration ; */

%macro prep_data(Input_data);
data pumf;
       set &Input_data;

       /* groupes d'âge tels que définis dans l'Annexe A du guide de l'utilisateur du
LFS PUMF ; */
                  if age_6 = '1'             then age_cal = '1516';
              else if age_6 = '2'            then age_cal = '1719';
              else if age_6 in ('3', '4')    then age_cal = '2024';
              else if age_12 = '03'          then age_cal = '2529';
              else if age_12 = '04'          then age_cal = '3034';
              else if age_12 in ('05', '06') then age_cal = '3544';
              else if age_12 in ('07', '08') then age_cal = '4554';
              else if age_12 = '09'          then age_cal = '5559';
              else if age_12 = '10'          then age_cal = '6064';
              else if age_12 = '11'          then age_cal = '6569';
              else if age_12 = '12'          then age_cal = '70+';
run;
%mend;
%prep_data(Input_data);


/* Calculer les facteurs d'ajustement pour chaque réplique bootstrap ; */
%macro calc_adj_fact(input);

  data adj_fact;
    set &input.;
    array pois_fact{&b.} pois_fact1 - pois_fact&b.;
    array adj_fact{&b.} adj_fact1 - adj_fact&b.;
       /* crée les facteurs d'ajustement &b. ; */
       do i = 1 to &b.;
           pois_fact[i] = 2 * (ranuni(&seed.) >= 0.5) - 1;
                     /* = 1 ou -1 avec une probabilité de 50 % ; */
            adj_fact[i] = 1 + pois_fact[i] * sqrt(1 - (1 / finalwt));
       end;
    drop i pois_fact:; /* Plus nécessaire ; */
   run;

%mend;
%calc_adj_fact(poisson_factors);
```

```
/* Calculer les poids bootstrap ; */

%macro generate_reps(input);
data uncal_bsw;
       set &input;

       array adj_fact{&b.} adj_fact1 - adj_fact&b.;
       /* Utiliser les facteurs d'ajustement pour créer les poids bootstrap ; */
       array bwun{&b.} bwun1 - bwun&b.;
       /* Crée les poids bootstrap &b. (non calibrés) ; */
              do i = 1 to &b.;
                     bwun[i] = finalwt * adj_fact[i];
              end;
       drop i adj_fact:; /* Plus nécessaire ; */
run;
%mend;
%generate_reps(adj_fact);


/* Obtenir les sommes des poids par province, groupe d'âge_cal et sexe pour la
calibration ; */
%macro calculate_sums(input);
proc summary data = &input;
       var finalwt;
       class prov age_cal gender;
       types prov*age_cal*gender;
       output out = totals_finalwt(drop = _TYPE_ _FREQ_) sum = sum_finalwt;
run;

proc summary data = &input;
       var bwun1 - bwun&b.;
       class prov age_cal gender;
       types prov*age_cal*gender;
       output out = totals_boots(drop = _TYPE_ _FREQ_)
           sum = sum_boot1-sum_boot&b.;
run;
%mend;
%calculate_sums(uncal_bsw);

/* Enfin, calibrez les bwun aux sommes des finalwts ; */
%macro calibrate_weights();
/* Fusionner les tables contenant les poids bootstrap non calibrés et les sommes
   que nous venons de calculer ; */

proc sort data = uncal_bsw; by prov age_cal gender; run;
data to_calibrate;
       merge uncal_bsw totals_finalwt totals_boots;
       by prov age_cal gender;
run;

/* Multiplier chaque bwun par la somme des finalwts / somme des bwuns ; */

data final_bs;
       set to_calibrate;

        array bw{&b.} bw1 - bw&b.;
              array bwun{&b.} bwun1 - bwun&b.;
              array sum_boot{&b.} sum_boot1 - sum_boot&b.;

           do i = 1 to &b.;
              bw[i] = bwun[i] * (sum_finalwt / sum_boot[i] );
```

```
          end;

             drop i bwun: sum_:;
run;
/* Laisser le jeu de données avec uniquement les poids bootstrap finaux ; */
%mend;


%calibrate_weights();

/* Le jeu de données final est "final_bs", qui contient toutes les variables PUMF et
les
   poids bootstrap calibrés. Vous pouvez ensuite utiliser ce jeu de données pour
calculer
   les estimations de variance comme bon vous semble ; */

/* Vous pouvez également combiner toutes les étapes dans une seule macro et tout
exécuter
   en même temps ; */

%macro generate_bootstrap_weights(Input_data);

%prep_data(Input_data);

*Combine les étapes 1-2;
data uncal_bsw;
set pumf;
    array adj_fact{&b.} adj_fact1 - adj_fact&b.;
    array bwun{&b.} bwun1 - bwun&b.;
           do i = 1 to &b.;
              adj_fact[i] = 2 * (ranuni(&seed.) >= 0.5) - 1;
              bwun[i] = finalwt * (1 + adj_fact[i] * sqrt(1 - (1 / finalwt)));
           end;
       drop i adj_fact1 - adj_fact&b.;
run;

%calculate_sums(uncal_bsw);
%calibrate_weights();

%mend;

%generate_bootstrap_weights(Input_data);

/*---------FIN DE LA CRÉATION DES POIDS BOOTSTRAP POISSON-----------------*/

/*---------EXEMPLES D'UTILISATION DES POIDS BOOTSTRAP POUR CALCULER LA VARIANCE-------
-*/
/* Chaque exemple peut prendre au moins quelques minutes                      */

/* Variance pour les totaux */
/* Exemple : total des chômeurs par province */

proc surveyfreq data=final varmethod=bootstrap; *spécifier bootstrap;
  tables prov*lfsstat /CLWT varWT CVWT nopercent nototal;
  weight finalwt;
  repweight bw1-bw&b.; /* Spécifier le nom des poids bootstrap créés ci-dessus ; */
  where lfsstat = '3'; *seulement calculer le total pour les chômeurs;
  ods output CrossTabs=Results_totals;
run;


/* Variance des taux/ratios */
/* Exemple : Taux de chômage par province */
```

```
* metre en place une variable indicatrice;
data pumf_unemp;
  set final;
          if lfsstat in ('1','2') then unemployed=0;
          else if lfsstat = '3'   then unemployed=1;
run;

proc surveyfreq data=pumf_unemp varmethod=bootstrap;
  tables prov*unemployed / nofreq oneway CL CV;
  weight finalwt;
  repweight bw1-bw&b.;
  where lfsstat in ('1','2','3'); *Taux de chômage est calculé uniquement auprès de la
population active;
run;

proc surveyfreq data=pumf_unemp varmethod=bootstrap;
  tables unemployed / nofreq oneway CL CV;
  weight finalwt;
  repweight bw1-bw&b.;
  where lfsstat in ('1','2','3'); *Taux de chômage est calculé uniquement auprès de la
population active;
by prov;
run;


/* Variance des moyennes pour les variables numériques */ /* Exemple : Revenus
horaires par province */

proc surveymeans data=pumf_results varmethod=bootstrap CV;
  var hrlyearn;
  weight finalwt;
  repweight bw1-bw&b.;
  by prov;
  where cowmain in ('1','2') and lfsstat in ('1','2'); *hrlyearn est uniquement
disponible pour les employés
run;
```

D2 : Exemple de programme du logiciel R

```
#===============================================================================

# CODE EXEMPLE POUR LE BOOTSTRAP POISSON SUR LE LFS PUMF

# Ce programme crée un fichier de 1 000 poids bootstrap poisson à utiliser

# avec le LFS PUMF, et montre également quelques exemples de la manière de

# calculer la variance de certaines estimations

# Input_data = un fichier LFS PUMF déjà chargé dans l'environnement

# seed = numéro afin de pouvoir reproduire les résultats aléatoires

# Packages requis :

# dplyr

# tidyr

#===============================================================================

# Charger les librairies
library(dplyr)
library(tidyr)

# Définir le seed et le nombre de répliques
seed = 1234
reps = 10 # Nombre de répliques. Utiliser 10 pour les tests et l'apprentissage. Utiliser 1000
pour la production.

# Effectué dans une série de function / Exécuter en bas pour suivre

# Comme étape préalable à la calibration des poids bootstrap, définissez les
# groupe de calibration comme indiqué dans l'annexe A du guide.
## Ajustez les niveau 1 et 2 de la variable age_6 comme niveau 0 et 1 de la variable age_12
## Séparer les groupes d'âge 15-19 en 15-16 et 17-19.
## convertir en "factor", fusionner les niveaux pour obtenir des catégorie de 10 ans
#  d'invervalle pour les 35-44 et 45-54.

prep_data <- function(pumf) {
  pumf$age_cal <- factor(
    ifelse(
      pumf$age_6 %in% 1:2,
      as.numeric(pumf$age_6)-1,
      as.numeric(pumf$age_12)
    ),
    levels = 0:12,
    labels = c(
      "15-16", "17-19", "20-24", "25-29", "30-34",
      rep("35-44", 2), rep("45-54", 2),
      "55-59", "60-64", "65-69", "70+"
    )
  )
  pumf
}

pumf <- prep_data(Input_data)
```

```r
# Fonction pour créer le facteur de poisson pour un individu
sample_poisson_factors <- function(input) {
  sample(c(-1, 1), length(input), replace = TRUE) # random and independent
}

# Example
# Crée une liste de 1 ou -1 pour chaque individus.
poisson_factors <- sample_poisson_factors(pumf$finalwt)


# function pour  calculer le poids boostrap non calibré
# (utilise la fonction sample_poisson_factors)
generate_replicates <- function(final_weight, n_reps, seed_value) {
  adjustment_factors <- final_weight * sqrt((final_weight - 1) / final_weight) # equation (1)
  set.seed(seed_value)

  replicate(
    n_reps,
    final_weight + sample_poisson_factors(final_weight) * adjustment_factors,   # equation (1)
+ (2)
    simplify = "array"
  ) |> as.matrix()
}

# exemple
uncal_bsw <- generate_replicates(pumf$finalwt, 10, seed)

# function pour calibrer chaque poids bootstrap en function du domaine
calibrate_weights <- function(uncalibrated_weights, final_weight, domains) {
  domain_indices <- split(seq_len(nrow(uncalibrated_weights)), domains)  #endroit ou retrouver
les group dans le FMGD

  domain_fw_totals <- domain_indices |>   # extraire ceux qui sont dans un group et calculer l
a somme des poids d'enquête
    sapply(function(x) sum(final_weight[x]))

  domain_bs_totals <- domain_indices |>          # extraire ceux qui sont dans un group et cal
culer la somme des poids d'enquête
    sapply(function(x) colSums(uncalibrated_weights[x, ]))

  domain_scaling_factors <- domain_fw_totals / t(domain_bs_totals)  # transposer la matrice

  uncalibrated_weights * domain_scaling_factors[domains, ]
}

# exemple
cal_bsw <- calibrate_weights(uncal_bsw, pumf$finalwt, interaction(pumf$prov, pumf$gender, pumf
$age_cal))


# Fonction finale qui combine les étapes précédentes
generate_bootstrap_weights <- function(d, n_reps, seed_value) {
  uncalibrated_weights <- generate_replicates(d$finalwt, n_reps, seed_value)
  domains <- interaction(d$prov, d$gender, d$age_cal)
  calibrate_weights(uncalibrated_weights, d$finalwt, domains)
}

# exemple
final_bs <- pumf |>
  dplyr::mutate(bswt = generate_bootstrap_weights(d=pumf, n_reps=reps, seed_value=seed))
```

```r
################################################################

#Example of using bootstrap weights to calculate variance

# define indicators
final_bs$employed <- final_bs$lfsstat %in% 1:2
final_bs$unemployed <- final_bs$lfsstat %in% 3
final_bs$nilf <- final_bs$lfsstat %in% 4

# Estimates of totals function
bs_total <- function(bootstrap_weights, final_weights) {
  est_fw <- sum(final_weights)
  est_bs <- colSums(bootstrap_weights)
  bs_var <- mean((est_bs - est_fw)^2)
  bs_sd <- sqrt(bs_var)
  bs_cv <- ifelse(est_fw != 0, abs(bs_sd / est_fw), 0)
  data.frame(
    est = est_fw,
    var = bs_var,
    sd = bs_sd,
    cv = bs_cv * 100,
    lb = est_fw - qnorm(0.975) * bs_sd,
    ub = est_fw + qnorm(0.975) * bs_sd,
    lbq = quantile(est_bs, 0.025),
    ubq = quantile(est_bs, 0.975),
    lbq2 = quantile(est_bs, 0.025, type=2),
    ubq2 = quantile(est_bs, 0.975, type=2)

  )
}

# Example de calcul pour le taux de chômage
unemp_by_prov <- function(bw_file) {
  res <- bw_file |>
    dplyr::filter(unemployed) |>
    dplyr::group_by(prov) |>
    dplyr::summarize(
      est = bs_total(bswt, finalwt),
      .groups = "drop"
    ) |>
    tidyr::unpack(cols = est, names_sep = "_")
  res
}

results <- unemp_by_prov(final_bs)

# Estimation de taux et de proportion
bs_ratio <- function(bootstrap_weights, final_weights, num) {
  final_weights_num = case_when(num==TRUE ~ final_weights, TRUE ~ 0)
  est_fw_num <- sum(final_weights_num)
  est_fw_den <- sum(final_weights)
  est_fw <- est_fw_num / est_fw_den
  bootstrap_weights_num = case_when(num==TRUE ~ bootstrap_weights, TRUE ~ 0)
  est_bs_num <- colSums(bootstrap_weights_num)
  est_bs_den <- colSums(bootstrap_weights)
  est_bs <- est_bs_num/est_bs_den
  bs_var <- mean((est_bs - est_fw)^2)
  bs_sd <- sqrt(bs_var)
  bs_cv <- ifelse(est_fw != 0, abs(bs_sd / est_fw), 0)
  data.frame(
```

```
    est = est_fw,
    var = bs_var,
    sd = bs_sd,
    cv = bs_cv * 100,
    lb = est_fw - qnorm(0.975) * bs_sd,
    ub = est_fw + qnorm(0.975) * bs_sd,
    lbq = quantile(est_bs, 0.025),
    ubq = quantile(est_bs, 0.975),
    lbq2 = quantile(est_bs, 0.025, type=2),
    ubq2 = quantile(est_bs, 0.975, type=2)

  )
}

# Exemple pour le taux de chômage par province
unemprate_by_prov <- function(bw_file) {
  res <- bw_file |>
    dplyr::filter(!nilf) |>  #calculate unemployment rate on those in labour force
    dplyr::group_by(prov) |>
    dplyr::summarize(
      est = bs_ratio(bswt, finalwt, unemployed),
      .groups = "drop"
    ) |>
    tidyr::unpack(cols = est, names_sep = "_")
  res
}

results_ratio <- unemprate_by_prov(final_bs)
```

Références

Beaumont, J.-F., & Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80(1), 127-148.

Statistics Canada. (2017, 12 21). *Methodology of the Canadian Labour Force Survey.* Retrieved from Statistics Canada: https://www150.statcan.gc.ca/n1/pub/71-526-x/71-526-x2017001-eng.htm