Guide abrégé de l’utilisateur des microdonnées

Enquête sur les modes d'apprentissage et de garde des jeunes enfants

Fichier de microdonnées à grande diffusion

2022

Statistics Canada Statistique Canada

Canada

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

Statistique Canada
Services à la clientèle
Centre canadien de la statistique de l'éducation (CCSE)
Téléphone : 613-951-3321 ou appelez sans frais 1-800-461-9050
Courriel : statcan.education-education.statcan@statcan.gc.ca

Accéder aux renseignements et passer une commande
L'Enquête sur les modes d'apprentissage et de garde des jeunes enfants (EMAGJE) produit deux types de fichiers de microdonnées : le fichier maître et le fichier de microdonnées à grande diffusion (FMGD). Pour plus de renseignements, consultez la page web suivante :
http://www.statcan.gc.ca/fr/microdonnees/centres-donnees

Fichiers maîtres
Les fichiers maîtres contiennent toutes les variables et tous les enregistrements recueillis lors de l'enquête et de sources administratives. Ces fichiers sont accessibles, moyennant un coût, à Statistique Canada pour un usage interne et dans les centres de données de recherche (CDR) de Statistique Canada, par l'intermédiaire de l'accès à distance en temps réel (ADTR) et dans le laboratoire virtuel de données (LVD). Les fichiers maîtres peuvent aussi faire l'objet de demandes de totalisations personnalisées.

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

Laboratoire virtuel de données
Le LVD est un centre d'accès infonuagique aux microdonnées qui permet aux chercheurs autorisés d'accéder à distance aux données hébergées par Statistique Canada au moyen d'une connexion informatique sécurisée et d'un environnement infonuagique protégé.

Totalisations personnalisées
Une autre méthode d'accès aux fichiers maîtres consiste à offrir à tous les utilisateurs de faire appel au personnel du service à la clientèle du Centre de l'intégration et du développement des données sociales pour produire des totalisations personnalisées. Ce service est offert moyennant le recouvrement des coûts. Il permet aux utilisateurs qui ne savent pas se servir de logiciels de totalisation d'obtenir des résultats personnalisés. Les résultats sont filtrés pour s'assurer qu'ils sont conformes aux normes de confidentialité et de fiabilité avant d'être diffusés. Pour plus de renseignements, communiquez avec le service à la clientèle au csdid-info-cidds@statcan.gc.ca.

Fichiers de microdonnées à grande diffusion
Les fichiers de microdonnées à grande diffusion sont élaborés à partir des fichiers maîtres suivant une technique qui vise à concilier l'impératif d'assurer la confidentialité des répondants et la nécessité de produire des données d'utilité maximale. Les FMGD doivent répondre à des normes sévères de sécurité et de confidentialité, conformément à la Loi sur la statistique avant qu'ils ne soient diffusés pour l'accès public. Pour s'assurer du respect de ces normes, chaque FMGD est soumis à un processus officiel d'examen et d'approbation par un comité formé de haut gestionnaire de Statistique Canada. Les variables les plus susceptibles de permettre l'identification d'une personne sont supprimées du fichier ou agrégées en catégories moins détaillées. Pour commander le FMGD, communiquez avec l'équipe Infostats au infostats@statcan.gc.ca.

Initiative de démocratisation des données

L'Initiative de démocratisation des données (IDD) permet aux étudiants et chercheurs d'utiliser les fichiers de microdonnées à grande diffusion dans plusieurs universités au Canada. Pour obtenir plus de renseignements, consultez la page web suivante : http://www.statcan.gc.ca/dli-ild/dli-idd-fra.htm

Table des matières

1.0 Introduction 5
2.0 Estimation 5
2.1 Lignes directrices pour la pondération 5
2.2 Lignes directrices pour l'arrondissement d'estimations 5
3.0 Qualité des données 6
3.1 Erreurs non dues à l'échantillonnage 6
3.2 Erreur d'échantillonnage 6
3.3 Lignes directrices relatives à la qualité pour la diffusion 7
3.4 Lignes directrices pour l'analyse statistique, l'estimation de la variance et la construction des intervalles de confiance 10

1.0 Introduction

Ce document fournit de l'information de base sur la production d'estimations pour le fichier de microdonnées à grande diffusion (FMGD) de L'Enquête sur les modes d'apprentissage et de garde des jeunes enfants (EMAGJE) et sur la diffusion de la qualité des estimations. Pour de l'information générale à propos de l'enquête, veuillez vous référer à l'information en ligne (IMDB).

2.0 Estimation

2.1 Lignes directrices pour la pondération

Le principe qui sous-tend une estimation pour un échantillon probabiliste veut que chacune des unités sélectionnées dans l'échantillon « représente », en plus d'elle-même, d'autres unités non-sélectionnées dans l'échantillon. Par exemple, si un échantillon aléatoire simple de 100 unités est sélectionné d'une population de 5 000 unités, chaque unité de l'échantillon représente 50 unités de la population. Le nombre d'unités que représente une unité de l'échantillon est le poids d'enquête de l'unité échantillonnée.

L'EMAGJE utilise un plan d'échantillonnage et une méthode d'estimation complexes, ce qui implique que les poids d'enquête ne sont pas égaux pour toutes les unités de l'échantillon. Lorsque des estimations et tableaux statistiques sont produits, les utilisateurs **doivent** appliquer le poids d'enquête approprié. Si les poids appropriés ne sont pas utilisés, 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.

2.2 Lignes directrices pour l'arrondissement d'estimations

Nous recommandons fortement aux utilisateurs de respecter les lignes directrices qui suivent pour l'arrondissement des estimations calculées à partir des fichiers de microdonnées :

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

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

c) Dans les cas, où, en raison de limitations d'ordre techniques ou de toutes autres limites, une technique d'arrondissement autre que la technique normale est utilisée pour produire des estimations à être publiées ou autrement diffusées différentes des estimations correspondantes publiées par Statistique Canada, nous recommandons fortement aux utilisateurs d'indiquer la raison de ces différences dans le ou les documents à publier ou à diffuser.

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

3.0 Qualité des données

La présente section traite des erreurs d'enquête (erreur non due à l'échantillonnage et erreur d'échantillonnage), des règles à utiliser pour déterminer si une estimation peut être diffusée pour des raisons de qualité, et du calcul des estimations de la variance et des intervalles de confiance.

3.1 Erreurs non dues à l'échantillonnage

Les erreurs d'enquête proviennent de diverses sources. Ces erreurs peuvent être classées en deux principales catégories : erreurs non dues à l'échantillonnage et erreurs d'échantillonnage. Les erreurs non dues à l'échantillonnage peuvent apparaître lors de presque toute activité d'enquête. Il se peut que les intervieweurs comprennent mal les instructions, que les répondants se trompent en répondant aux questions, que des réponses soient mal entrées dans le système et que des erreurs se produisent au moment du traitement et de la tabulation des données. Ces erreurs sont des exemples d'erreur non due à l'échantillonnage. Si l'on considère un grand nombre d'observations, les erreurs aléatoires ont peu d'effets sur les estimations tirées des données de l'enquête. Toutefois les erreurs qui se produisent systématiquement peuvent causer des biais dans les estimations.

La non-réponse est une source à la fois d'erreur non due à l'échantillonnage et d'erreur d'échantillonnage. La non-réponse résulte de l'inhabilité de recueillir de l'information complète pour toutes les unités sélectionnées dans l'échantillon. La non-réponse est une source d'erreur non due à l'échantillonnage dans le sens où les non-répondants ont souvent des caractéristiques différentes de celles des répondants. Cela peut mener à des estimations de l'enquête biaisées si le biais de la non-réponse n'est pas entièrement éliminé via des ajustements aux poids. Plus bas est le taux de réponse, plus important est le risque de biais. La non-réponse est aussi une source d'erreur d'échantillonnage ; ceci est discuté davantage dans la Section 3.2. Un taux de réponse de 58,0% a été obtenu pour L'EMAGJE.

3.2 Erreur d'échantillonnage

L'erreur d'échantillonnage est définie comme étant l'erreur résultant de l'estimation d'une caractéristique de la population du fait de mesurer une partie au lieu de toute la population. Des méthodes de calcul de l'erreur d'échantillonnage s'appliquent à l'enquête-échantillon probabiliste. Ces méthodes découlent directement du plan d'échantillonnage et de la méthode d'estimation de l'enquête.

La mesure utilisée le plus souvent pour quantifier l'erreur d'échantillonnage est la variance d'échantillonnage. La variance d'échantillonnage mesure à quel point l'estimation d'une caractéristique à partir de divers échantillons possibles de même taille et de même plan de sondage est différente l'une de l'autre. Dans le cas des plans d'échantillonnage qui utilisent l'échantillonnage probabiliste, l'amplitude de la variance d'échantillonnage d'une estimation peut être estimée.

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

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

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

Le coefficient de variation (CV) d'une estimation est une mesure relative de l'erreur d'échantillonnage. Il est défini comme l'estimation de l'erreur-type divisée par l'estimation elle-même. Il est très utile pour mesurer et comparer l'erreur d'échantillonnage de variables quantitatives avec de grandes valeurs positives. Cependant, il n'est pas recommandé pour des estimations telles que les proportions, les estimations 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'échantillonnage 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), l'intervalle de confiance couvrirait la véritable valeur dans la population.

3.3 Lignes directrices relatives à la qualité pour la diffusion

Avant de diffuser et/ou de publier toute estimation, les analystes doivent prendre en considération le niveau de qualité de l'estimation. Tandis que les erreurs d'échantillonnage et non dues à l'échantillonnage influencent la qualité des données, la présente section porte sur la qualité en ce qui a trait aux erreurs d'échantillonnage. À Statistique Canada, il est considéré comme une pratique exemplaire de faire état de l'erreur d'échantillonnage d'une estimation par l'intermédiaire de son intervalle de confiance (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
Les estimations et intervalles de confiance peuvent être publiés sans mise en garde. Les utilisateurs des données devraient utiliser l'intervalle de confiance de 95 % pour déterminer si la qualité de l'estimation est suffisante. Notez que le « A » n'est pas un indicateur de qualité; il ne devrait pas être diffusé.

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

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

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

Règles de diffusion pour les estimations de proportions et de comptes

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

Les règles de diffusion pour les estimations de proportions et de comptes sont basées sur la taille d'échantillon. Les tableaux 1 et 2 fournissent les règles de diffusion pour le FMGD de l'EMAGJE. Les règles du tableau 1 sont utilisées lorsque le domaine d'intérêt est au niveau du Canada ou au niveau du Canada par group d'âge. Les règles du tableau 2 sont utilisées lorsque le domaine d'intérêt est au niveau provincial ou à un niveau inférieur; en d'autres termes, tous les répondants qui contribuent à l'estimation appartiennent à la même province. Cela inclut le niveau de la province par groupe d'âge.

Tableau 1 : Règles générales pour les proportions et les comptes

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n ≥ 115	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
60 ≤ n < 115	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
n < 60	F	Supprimer l'estimation et son IC pour des raisons de qualité

Tableau 2 : Règles pour les proportions et les comptes pour des estimations au niveau provincial ou inférieur

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n ≥ 80	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
40 ≤ n < 80	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
n < 40	F	Supprimer l'estimation et son IC pour des raisons de qualité

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

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

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

Les règles de diffusion pour des estimations de moyennes et totaux de variables quantitatives sont basées sur la taille d'échantillon et sur le CV de l'estimation. Les tableaux 3 et 4 fournissent les règles de diffusion pour le FMGD de l'EMAGJE. Les règles du tableau 3 sont utilisées lorsque le domaine d'intérêt est au niveau du Canada ou au niveau du Canada par group d'âge. Les règles du tableau 4 sont utilisées lorsque le domaine d'intérêt est au niveau provincial ou à un niveau inférieur; en d'autres termes, tous les répondants qui contribuent à l'estimation appartiennent à la même province. Cela inclut le niveau de la province par groupe d'âge.

Tableau 3 : Règles générales pour les moyennes et totaux

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n≥115 et CV≤25 %	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
Autrement	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
n<60 ou CV>50 %	F	Supprimer l'estimation et son IC pour des raisons de qualité

Tableau 4 : Règles générales pour les moyennes et totaux des estimations au niveau provincial ou inférieur

Taille d'échantillon (n)	Catégorie de Diffusion	Action
n≥80 et CV≤25 %	A*	Diffuser sans mise en garde; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
Autrement	E	Diffuser avec une mise en garde sur la qualité; les utilisateurs devraient utiliser l'IC comme indicateur de qualité
n<40 ou CV>50 %	F	Supprimer l'estimation et son IC pour des raisons de qualité

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

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

Règles de diffusion pour les différences

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

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

Règles additionnelles au sujet des intervalles de confiance

Les règles de diffusion précédentes devraient supprimer la plupart des estimations et intervalles de confiance de piètre qualité. Il y a également deux conditions qui indiquent si un intervalle de confiance est de piètre qualité. Une estimation et son intervalle de confiance devraient être assignés à la catégorie de diffusion F si l'une ou l'autre des deux conditions suivantes est satisfaite :

- La borne inférieure de l'intervalle de confiance de 95 % est égale à la borne supérieure de l'intervalle; en d'autres termes, l'intervalle de confiance est de longueur nulle. (Une exception survient si l'estimation correspond à un total de contrôle 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.

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

Progiciels statistiques pour l'analyse statistique et l'estimation de la variance

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

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

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

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

Intervalles de confiance

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

Les exemples ci-dessous montrent comment ces méthodes recommandées pour contruire des intervalles de confiance sont précisées pour des proportions dans SAS et SUDAAN.

1. SAS, intervalles de confiance de Wilson modifiés :

```
PROC SURVEYFREQ
DATA=…. VARMETHOD=BRR;
WEIGHT WTMP;
REPWEIGHTS WRMP1-WRMP1000;
TABLES .... / CL (TYPE=WILSON ADJUST=NO TRUNCATE=YES)
```

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

```
PROC CROSSTAB
DATA=…. DESIGN=BRR SMCONF=50;
WEIGHT WTMP;
REPWGT WRMP1-WRMP1000;
TABLES ...;
```

Rééchelonnage des poids

Comme nous l'avons déjà mentionné, il est recommandé que les utilisateurs utilisent les procédures d'analyse conçues pour analyser des données à partir de plans d'échantillonnage complexes, qui peuvent utiliser les poids pour produire des estimations et les poids bootstrap pour produire des estimations de la variance. Certaines procédures d'analyse qui ne sont pas conçues pour le contexte de l'échantillonnage peuvent permettre l'utilisation des poids (sans poids bootstrap). Cependant, ces procédures peuvent différer dans leur définition des poids, et produire des estimations exactes mais des estimations de la variance dénuées de sens. Pour des analyses telles que la régression linéaire, la régression logistique et l'analyse de variance, le rééchelonnement des poids peut rendre la variance calculée par les progiciels standards plus raisonnable. Les poids pour le domaine d'intérêt devraient être rééchelonnés de sorte que le poids moyen est d'un (1); ceci peut être accompli en divisant chaque poids par le poids moyen global du domaine avant de procéder à l'analyse. Le rééchelonnage rend les estimations de la variance plus raisonnables, mais celles-ci tiennent compte uniquement de l'inégalité des probabilités de sélection – elles ne tiennent pas compte ni de la stratification ni des grappes du plan d'échantillonnage. Cette approche devrait donc être utilisée qu'en dernier recours, seulement lorsque qu'aucune procédure permettant l'utilisation des poids bootstrap n'est disponible; les utilisateurs sont avisés que les résultats sont approximatifs.

Références

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

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

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