Nº 12-002-XIF au catalogue

Le Bulletin technique et d'information des Centres de données de recherche

Printemps 2005, vol. 2 nº 1

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Le programme des Centres de Données de Recherche, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-002-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. 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 dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada
Le programme des centres de données de recherche

Le Bulletin technique et d'information des Centres de données de recherche

Printemps 2005, vol. 2 n° 1

Juin 2005

N° 12-002-XIF au catalogue

Périodicité: semestriel

ISSN : 1710-2200

Ottawa

This publication is available in English (Catalogue no. 12-002-XIE)

Note de reconnaissance

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

À propos du Bulletin technique et d'information

Le Bulletin technique et d'information des Centres de données de recherche est un forum permettant aux utilisateurs actuels et prospectifs des centres de partager de l'information et les techniques d'analyse des données disponibles dans les centres. Le bulletin paraît au printemps et à l'automne, et l'on publiera à l'occasion des numéros spéciaux sur des questions d'actualité.

Objectifs

Les objectifs principaux de ce bulletin sont les suivants :

- l'accroissement et la diffusion de la connaissance concernant les données de Statistique Canada;
- les échanges d'idées parmi les utilisateurs membres des Centres de données de recherche (CDR);
- l'aide aux nouveaux utilisateurs du programme CDR; et
- offrir des occasions supplémentaires permettant aux chercheurs dans les centres de communiquer avec les spécialistes et divisions spécialisées au sein de Statistique Canada.

Le contenu

Nous souhaitons publier des articles qui contribueront à accroître la qualité des travaux de recherché menés dans les Centres de données de recherche de Statistique Canada et qui fourniront des conseils méthodologiques aux chercheurs travaillant dans les CDR.

Les articles figurant dans le Bulletin technique et d'information portent principalement sur :

- l'analyse et la modélisation des données;
- la gestion des données;
- les pratiques statistiques, informatiques ou scientifiques éprouvées ou au contraire inefficaces;
- le contenu en données;
- les effets associés au libellé des questionnaires;
- la comparaison d'ensembles de données;
- l'examen des méthodes et de leur application;
- les particularités que présentent les données;
- les problèmes associés aux données et leurs solutions; et
- les outils innovateurs faisant appel aux enquêtes et aux logiciels pertinents des CDR.

Ceux et celles qui s'intéressent à soumettre un article au Bulletin technique et d'information sont priés de suivre les directives pour les auteurs.

Les rédacteurs et les auteurs tiennent à remercier les réviseurs de leurs commentaires précieux.

Table des matières

Comparaison d'un taux dans une sous-population au taux dans l'ensemble de la population : Comment procéder à partir de données d'enquête et à l'aide des outils logiciels disponibles

Par Yves Lafortune et Georgia Roberts

Résumé

Il arrive souvent qu'on veuille utiliser des microdonnées d'enquête pour déterminer si le taux de fréquence d'une caractéristique donnée dans une sous-population est le même que celui dans l'ensemble de la population. Le présent document expose diverses façons de procéder pour faire des inférences au sujet d'une différence entre les taux et montre si et comment ces options peuvent être mises en œuvre dans trois différents progiciels d'enquête. Tous les progiciels illustrés, soit SUDAAN, WesVar et Bootvar, peuvent utiliser les poids bootstrap fournis par l'analyste pour procéder à l'estimation de la variance.

Introduction

On cherche souvent à savoir si le taux de fréquence d'une caractéristique donnée dans une sous-population est le même que celui dans l'ensemble de la population. Par exemple, une autorité sanitaire peut se demander si le taux d'incidence de la grippe dans sa région sociosanitaire est le même que le taux d'incidence pour la province dans son ensemble, ou bien une province peut vouloir savoir si la proportion d'élèves en 9^{e} année qui ont fait au moins dix heures de bénévolat au cours de la dernière année est la même que la proportion globale d'élèves du secondaire qui ont fait du bénévolat dans la même mesure. Les chercheurs veulent souvent examiner pareilles questions au moyen de données d'enquête. Nous expliquons ci-dessous comment procéder, d'abord en théorie, puis du point de vue pratique, lorsque les microdonnées d'enquête, les poids de rééchantillonnage et des progiciels d'enquête comme Bootvar (un ensemble de macros SAS ou SPSS), SUDAAN ou WesVar sont disponibles.

II. Un peu de théorie

Supposons qu'on souhaite comparer le taux d'usage du tabac en Ontario et le taux global d'usage du tabac au Canada. Soit p le véritable taux d'usage du tabac dans la population canadienne et soit p_{ONT} le véritable taux d'usage du tabac dans la province de l'Ontario. L'hypothèse à vérifier est alors $H_0 : p_{ONT} = p$. Pour vérifier cette hypothèse (comparativement à l'hypothèse alternative selon laquelle les deux proportions ne sont pas égales), on pourrait utiliser la statistique $T = (\hat{p}_{ONT} - \hat{p}) \Big/ \sqrt{\hat{\text{var}}(\hat{p}_{ONT} - \hat{p})}$, où $\hat{p}$, $\hat{p}_{ONT}$ et $\hat{\text{var}}(\hat{p}_{ONT} - \hat{p})$ sont des estimations établies à partir des données d'enquête. Pour vérifier l'hypothèse, on pourrait comparer cette statistique aux quantiles appropriés d'une distribution normale ou d'une distribution *t*, ou bien examiner la valeur p de la statistique. Par exemple, si le test de vérification est effectué au niveau de signification de 95 %, et si l'on considère que *T* suit une loi de distribution normale, on peut

déclarer que p diffère significativement de p_{ONT} si la valeur de T est supérieure à 1,96 ou inférieure à -1,96 ou si la valeur p donnée pour la statistique est inférieure à 0,05.

Il est facile d'obtenir les estimations $\hat{p}$ et $\hat{p}_{ONT}$ à l'aide du logiciel d'enquête, de même que les estimations de leurs variances individuelles. Toutefois, pour vérifier l'hypothèse d'intérêt, nous avons besoin d'une estimation de la variance de la différence, et c'est là que le problème se pose. Comme $\hat{p}$ et $\hat{p}_{ONT}$ ne sont pas estimés à partir d'échantillons indépendants (en fait, l'échantillon pour l'Ontario a été sélectionné dans le cadre de l'échantillon pour le Canada) la variance de la différence comprend une composante de covariance dont il faut tenir compte (rappelons que $\text{var}(\hat{p}_{ONT} - \hat{p}) = \text{var}(\hat{p}_{ONT}) + \text{var}(\hat{p}) - 2\,\text{cov}(\hat{p}_{ONT}, \hat{p})$). En fait, plus la taille de la sous-population est grande, plus il est probable que ce terme de covariance soit non négligeable.

Dans ces conditions, comment procéder pour obtenir une valeur appropriée pour le dénominateur de la variable à tester T pour l'hypothèse d'intérêt? Pourrait-il même y avoir une différente variable à tester pour la même hypothèse, que nous pourrions calculer à l'aide du logiciel disponible?

Solution n° 1 : Si une méthode d'estimation de la variance par rééchantillonnage, comme le « bootstrap », est recommandée pour l'enquête, et si des poids de rééchantillonnage sont disponibles aux fins d'estimation générale de la variance, alors on peut produire une estimation de $p_{ONT} - p$ en utilisant chacun de ces poids de rééchantillonnage et utiliser ensuite ces estimations répétées, de même que l'estimation calculée pour l'ensemble de l'échantillon, pour obtenir une estimation de la variance pour la différence estimée entre les proportions (la formule de calcul de l'estimation de la variance dépendant de la méthode de rééchantillonnage utilisée). Une personne pourrait rédiger son propre programme pour ce faire, ou choisir un outil logiciel qui produit ces estimations dans le cadre de ses sorties courantes. Il convient de souligner que, selon cette méthode, il n'est pas nécessaire de calculer explicitement les composantes de covariance de la variance de la différence.
[La solution n° 1 peut être appliquée au moyen de WesVar ou de Bootvar, mais non au moyen de SUDAAN.]

Solution n° 2 : Une autre possibilité consiste à utiliser un outil logiciel qui fournit une estimation de la matrice de covariance complète d'un ensemble de proportions (ou pourcentages) estimées, en version imprimée ou sous forme d'ensemble de données de sortie. Cette matrice de covariance doit contenir les estimations de la variance du taux estimatif pour l'ensemble de la population et du taux estimatif correspondant pour la sous-population. Elle doit contenir également une estimation de la covariance des deux taux estimatifs. À partir de ces quantités, on pourrait alors calculer une estimation de la variance de la différence entre les deux taux estimatifs, au moyen de la formule $\hat{\text{var}}(\hat{p}_{ONT} - \hat{p}) = \hat{\text{var}}(\hat{p}_{ONT}) + \hat{\text{var}}(\hat{p}) - 2\,\text{cov}(\hat{p}_{ONT}, \hat{p})$, et cette estimation pourrait alors servir à calculer la statistique T décrite ci-dessus. S'il n'est pas simple d'extraire électroniquement les quantités requises d'une matrice de covariance des sorties, il pourrait être plus facile d'utiliser papier et crayon pour calculer la statistique T requise. Il est peu probable qu'on choisisse la solution n° 2 si la solution n° 1 est facile à appliquer à l'aide de l'outil logiciel utilisé, puisque la solution n° 2 est plus laborieuse.

[La solution n^o 2 peut être appliquée au moyen de SUDAAN, mais non au moyen de WesVar ou de Bootvar.]

Solution n^o 3 : Les logiciels permettent souvent d'utiliser une variable à tester qui donne lieu à un contraste linéaire entre deux taux calculés pour deux sous-populations, mais non une variable à tester donnant lieu à un contraste linéaire entre un taux calculé pour une sous-population et un taux calculé pour une population. Par conséquent, il existe une façon différente de s'attaquer au problème, qui est la suivante. Il est facile de montrer (voir l'annexe 1) que l'hypothèse à vérifier, $H_0 : p_{ONT} = p$, est équivalente à l'hypothèse $H_0 : p_{ONT} = p_{ONT^c}$, où p_{ONT^c} est le véritable taux d'usage du tabac dans le reste du Canada (c.-à-d., dans l'ensemble de la population dont la sous-population d'intérêt a été retranchée). Pour vérifier cette hypothèse, on peut utiliser la variable à tester $T_2 = (\hat{p}_{ONT} - \hat{p}_{ONT^c}) \Big/ \sqrt{\hat{\text{var}}(\hat{p}_{ONT} - \hat{p}_{ONT^c})}$. Cette variable à tester ne donne lieu à un contraste linéaire qu'entre deux taux de sous-populations. Veuillez noter que les deux variables à tester (T et T_2) ne sont pas les mêmes (bien que toutes deux fournissent une réponse à la même question), mais qu'elles sont fort susceptibles de mener à la même conclusion.
[La solution n^o 3 peut être appliquée au moyen de l'un quelconque des trois logiciels sélectionnés; toutefois, il convient le mieux à SUDAAN pour lequel la solution n^o 1 n'est pas disponible et la solution n^o 2 est difficile à exécuter complètement par voie électronique.]

Il convient de souligner que toutes trois solutions peuvent être appliquées tant dans le cas où la sous-population et le reste de la population complète sont échantillonnées indépendamment que dans celui où il y a dépendance entre les échantillons dans la sous-population et le reste de la population complète.

III. Mise en œuvre de ces solutions au moyen du logiciel d'enquête

Dans les exemples ci-dessous, nous utilisons le fichier de données **synthétiques** sur la santé (h356) du Cycle 3 (1998-1999) de l'Enquête nationale sur la santé de la population (ENSP) pour illustrer comment appliquer l'une ou plusieurs des solutions au moyen des différents logiciels. Les demandes de données doivent être adressées à « Unité de l'accès aux données, Enquête sur la santé de la population, Division de la statistique de la santé », courriel : nphs-ensp@statcan.ca et/ou cchs-escc@statcan.ca. Les renseignements tirés du plan de sondage de l'ENSP comprennent, pour chaque personne, un poids final (WT68), ainsi que 500 poids de rééchantillonnage **bootstrap** (BSW1-BSW500). L'identificateur unique pour chaque personne est donné par la combinaison des variables REALUKEY et PERSONID. Nous supposons que l'utilisateur a accès à un ensemble de données SAS qui contient les données de l'enquête. Pour chaque exemple, nous indiquons la façon de procéder pour préparer les données de la manière appropriée.

Le fichier de données synthétiques du Cycle 3 de l'ENSP (h356) comprend les réponses données par 17 244 personnes âgées de 0 à 99 ans. Les questions portant sur l'usage du tabac n'ont été posées qu'aux personnes de 12 ans ou plus. Par conséquent, les enregistrements de 1 995 personnes âgées de 0 à 11 ans ne sont pas pertinents aux fins de la présente analyse,

puisque les questions sur l'usage du tabac n'ont pas été posées à ces personnes. En outre, les réponses à la question d'intérêt portant sur l'usage du tabac manquent dans le cas de 32 personnes. Étant donné qu'il s'agit d'une petite proportion de l'échantillon, nous supposons que le fait de ne pas traiter la non-réponse n'a pas d'incidence sur les résultats.

Ainsi, les résultats devraient être fondés sur un nombre total de 15 217 personnes, représentant un total de 24 859 391 Canadiens âgés de 12 ans ou plus. Dans tous trois logiciels, nous utilisons le fichier complet contenant les données sur l'ensemble des 17 244 personnes, mais le codage et les méthodes utilisées pour obtenir les résultats garantissent que nos résultats sont fondés seulement sur les 15 217 répondants valides.

Nous utilisons les variables suivantes :

PRC8_CUR : « Province de résidence au moment de la collecte des données en 1998-1999 ». Les valeurs possibles sont T.-N. (10), Î.-P.-É. (11), N.-É. (12), N.-B. (13), Qué. (24), Ont. (35), Man. (46), Sask. (47), Alb. (48) et C.-B. (59).

SMC8_2 : « Actuellement, fumez-vous des cigarettes tous les jours, à l'occasion ou jamais? » Les valeurs possibles sont QUOTIDIENNEMENT (1), À L'OCCASION (2), JAMAIS (3), SANS OBJET (6), NE SAIS PAS (7), REFUS (8) et NON DÉCLARÉ (9). Nous définissons comme étant un « fumeur régulier » une personne qui fumait des cigarettes quotidiennement au moment de la collecte des données (c.-à-d., SMC8_2=1).

Nous remplaçons les valeurs 6, 7, 8 et 9 de la variable SMC8_2 par une valeur manquante « . » pour les 1 995 + 32 répondants non valides, de manière à garantir que les résultats obtenus au moyen de WesVar et de SUDAAN ne sont fondés que sur les 15 217 répondants valides. (Il convient de souligner qu'il serait possible d'obtenir les mêmes résultats tout en conservant les codes originaux, mais qu'il faudrait utiliser des options plus évoluées dans WesVar et SUDAAN.) Avec Bootvar, des étapes additionnelles sont requises pour s'assurer de la validité des résultats et nous les exposons à l'exemple 4.

Nous voulons donc déterminer le taux d'usage régulier du tabac chez les personnes de 12 ans et plus. Nous souhaitons comparer les taux d'usage du tabac en Ontario et dans l'ensemble du Canada.

IV. Exemple 1 – Utilisation de WesVar 4.2 pour illustrer la solution n^{o} 1 :

WesVar 4.2 est capable d'importer des fichiers de données dans la version 8 de SAS. Une fois les données disponibles dans WesVar, l'utilisateur doit attribuer chaque variable comme il se doit à son usage respectif. La case « Full Sample » doit contenir le poids final, la case « ID », les variables de l'identificateur unique, la case « Replicates », les variables qui sont les poids de rééchantillonnage et la case « Variables », toutes les variables d'intérêt restantes (qui peuvent être beaucoup plus nombreuses que celles nécessaires aux fins de cet exemple). Même si un plan d'enquête bootstrap n'est pas disponible directement dans WesVar, Phillips (2004) montre qu'il est possible d'utiliser l'option BRR dans WesVar pour calculer les estimations de la variance

bootstrap, à la condition que les poids bootstrap aient été produits à l'extérieur du logiciel. Par conséquent, la façon correcte de préparer les données synthétiques du Cycle 3 de l'ENSP dans WesVar est celle indiquée dans la figure ci-dessous (figure 1):

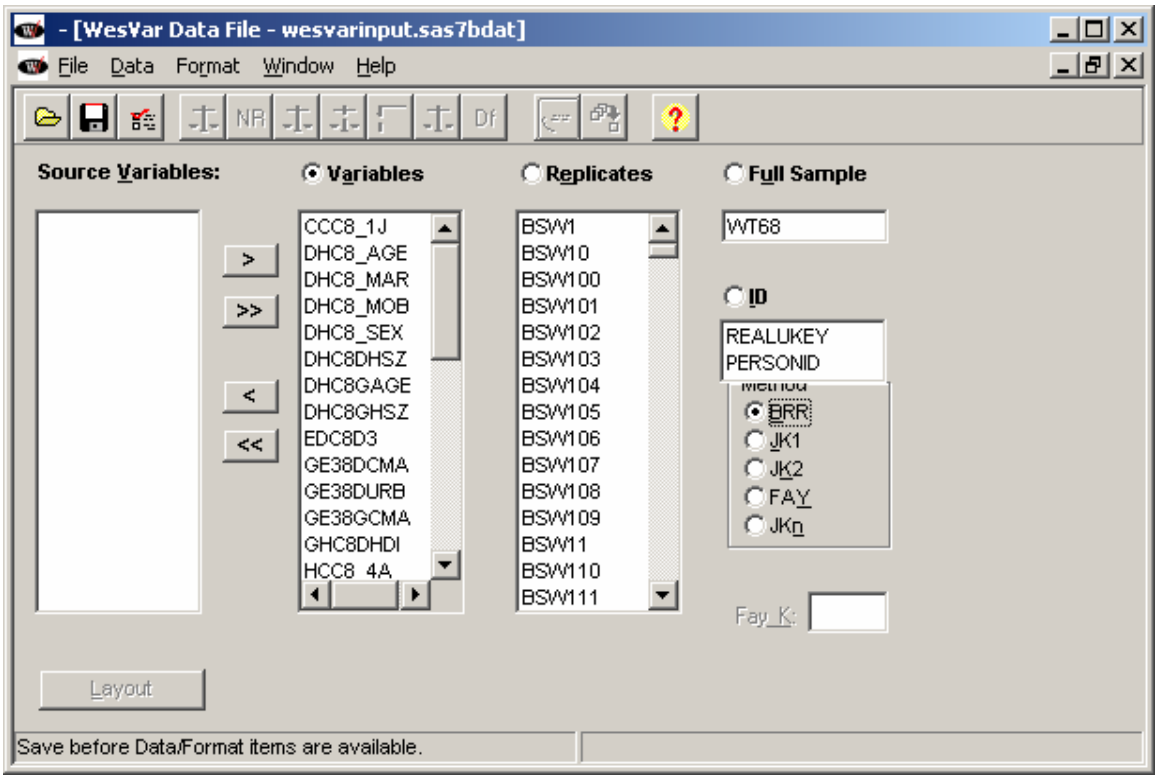

Figure 1

Nous pouvons ensuite poursuivre en sauvegardant d'abord les données et en sortant du tableau, puis en créant un nouveau cahier de travail. Nous rappelons que la variable identifiant la province de résidence est donnée par PRC8_CUR (avec une valeur de 35 pour l'Ontario), tandis que la variable identifiant les types de fumeurs est SMC8_2 (avec une valeur de 1 pour le fumeur régulier). Nous produirons maintenant les taux d'usage du tabac pour le Canada et l'Ontario. Pour ce faire, il faut présenter une nouvelle demande de tableau. D'abord, une autre statistique générée sera ajoutée puisque, par défaut, les valeurs p ne font pas partie de la sortie. Pour ce faire, on clique d'abord sur la zone Generated Statistics et l'on ajoute une coche dans la dernière case. Voici le tableau qui devrait s'afficher à l'écran (figure 2):

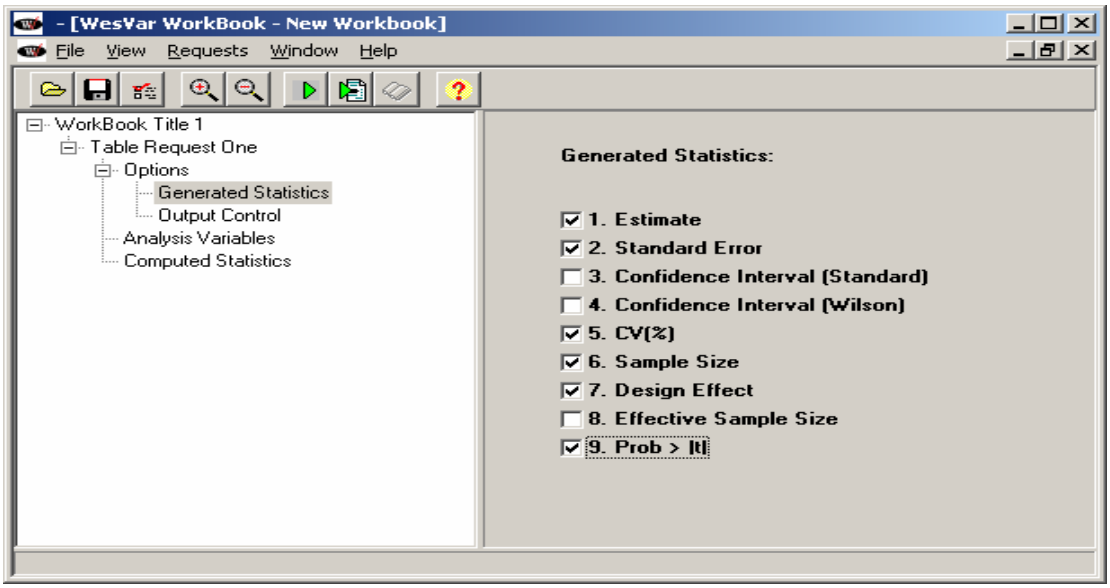

Figure 2

Ensuite, l'utilisateur peut cliquer sur le nœud Table Request au panneau de gauche de l'écran, puis sur Add Table Set (Single) et attribuer les variables requises à la case Selected (voir figure 3). Ici, les pourcentages d'intérêt sont ceux des lignes (row percents), puisque PRC8_CUR était la première variable sélectionnée.

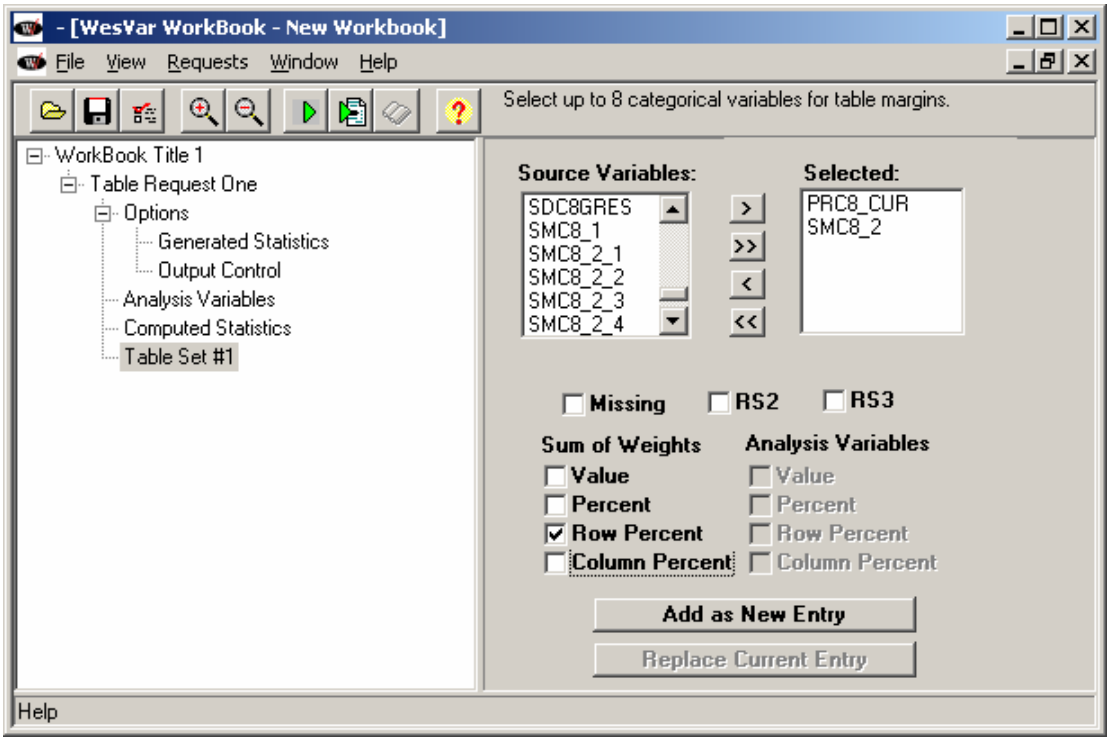

Figure 3

Ajoutez ce tableau comme nouvelle entrée et cliquez sur le petit signe + qui figure à gauche du nom du tableau (étiquette à côté des noms des variables qui constituent le tableau). Vous avez alors accès à trois autres panneaux (Cells, Cell Functions et Standardized Rates). Cliquez d'abord sur *Cells* pour attribuer des étiquettes aux cellules dans le tableau qui revêt un intérêt particulier. Par exemple, l'étiquette *Can_Smk_rate* sera attribuée à la cellule ayant une valeur marginale (« Marginal ») pour PRC8_CUR et de « 1 » pour SMC8_2, tandis que l'étiquette *Ont_Smk_rate* sera attribuée à la cellule ayant une valeur de « 35 » pour PRC8_CUR et de « 1 » pour SMC8_2, comme ci-dessous (figure 4).

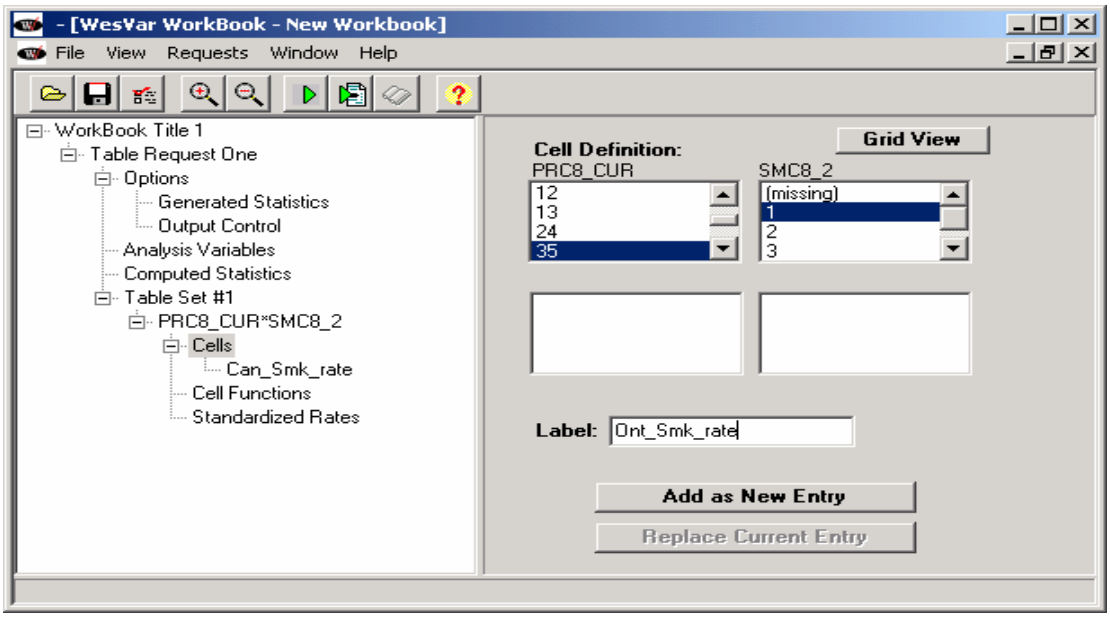

Figure 4

Ensuite, comme nous voulons obtenir la différence entre le taux d'usage du tabac pour l'Ontario et celui pour le Canada, nous utilisons également le panneau *Cell Functions*. La statistique « Diff_Smk_rate » = « Ont_Smk_rate » - « Can_Smk_rate » s'ajoute (voir figure 5).

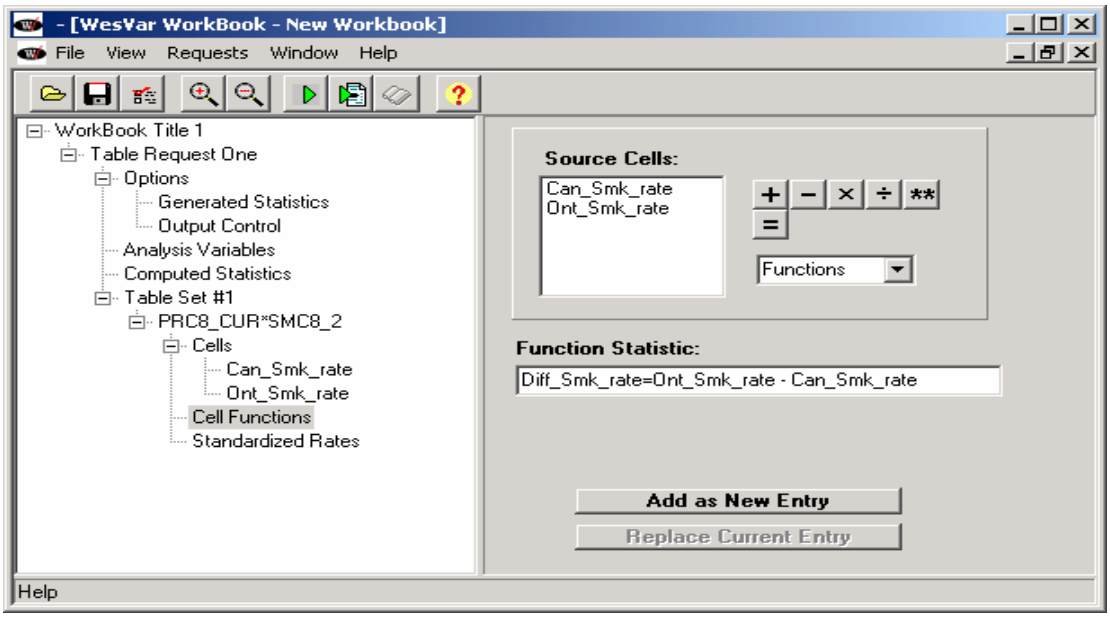

Figure 5

Exécutons maintenant le Table Request et examinons les résultats en cliquant sur l'icône représentant un livre ouvert (une fois disponible). Nous cliquons ensuite sur le nœud Functions dans le panneau de gauche pour examiner les résultats de la comparaison demandée, qui sont indiqués dans la figure ci-dessous (figure 6).

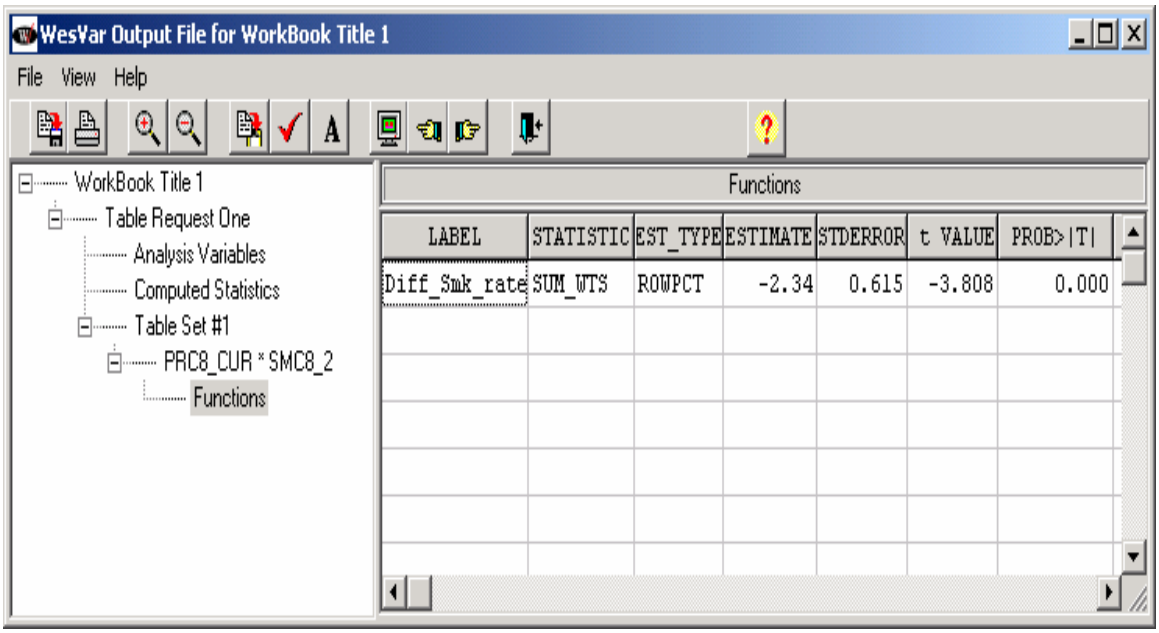

Figure 6

La différence estimée entre le taux d'usage du tabac pour l'Ontario et celui pour le Canada est d'environ 2,34 %. La valeur de la statistique T est appelée *t Value* dans la sortie et la valeur p est appelée PROB>|T|. Comme la valeur p pour Diff_Smk_rate est inférieure à 0,05, nous pouvons conclure, au niveau nominal de 5 %, à une différence significative entre le taux d'usage du tabac pour le Canada et celui pour l'Ontario. Il faut se rappeler que ces conclusions sont fondées sur des données synthétiques et qu'il ne faut donc absolument pas s'y fier.

V. Exemple 2 – Utilisation de SUDAAN 8.0.2 ou de SUDAAN 9.0.1 pour illustrer la solution n° 2 :

SUDAAN est maintenant appelable dans SAS, ce qui veut dire qu'il peut être utilisé dans un environnement SAS. Pour la plupart des usages, cela équivaut à la mise à la disposition de l'utilisateur de SAS de nouvelles procédures SAS additionnelles. Lorsqu'il aura appris les spécificités du codage des procédures SUDAAN, l'utilisateur pourra fonctionner aussi bien en SUDAAN qu'en SAS.

Avec les données de l'ENSP, pour calculer les estimations de la variance bootstrap, l'utilisateur doit procéder comme dans WesVar et préciser que le plan est un plan à répliques répétées équilibrées (BRR), mais inclure les poids bootstrap dans le fichier de données comme s'ils étaient des poids BRR.

PROC DESCRIPT est une procédure SUDAAN capable de calculer des totaux, des moyennes, des proportions et des percentiles pour des populations et des sous-populations. Elle a la capacité de créer dans le fichier de sortie la matrice de variance-covariance complète des estimations requises, ainsi que les covariances entre les estimations de la population complète et toutes estimations de sous-populations requises. Par conséquent, il est possible de mettre en application la solution n° 2. Voici une partie de code utilisant SUDAAN 8.0.2 PROC DESCRIPT qui permet à l'utilisateur de ce faire.

```
proc descript data=nphsdum3 design=brr;
   weight wt68;
   repwgt bsw1-bsw500;
   recode prc8_cur=(10 11 12 13 24 35 46 47 48 59);
   subgroup prc8_cur;
   levels 10;
   var smc8_2;
   catlevel 1;
   tables prc8_cur;
   output / pctcov=all filename=out1 REPLACE filetype=sas;
run;
```

Cette procédure calcule la proportion de fumeurs réguliers (SMC8_2 avec une valeur de 1) séparément pour toutes les provinces ainsi que pour l'ensemble du pays. En outre, elle crée un fichier de sortie appelé out101 (SUDAAN ajoute un suffixe) contenant, entre autres, les taux estimés (à la première ligne), ainsi que la matrice de variance-covariance de toutes les estimations (aux lignes subséquentes). Le tableau ci-dessous montre les valeurs pour le Canada et pour chaque province. Pour faciliter la consultation, les chiffres qui revêtent un intérêt particulier pour l'exemple sont indiqués en caractères gras (voir tableau 1).

Tableau 1

CANADA	T.-N.	Î.-P.-É.	N.-É.	N.-B.	Qué.	Ont.	Man.	Sask.	Alb.	C.-B.
24.241	24.921	26.444	26.615	26.988	28.163	**21.899**	24.710	25.340	25.144	21.445
0.218	0.092	-0.001	0.087	0.038	0.310	**0.207**	0.143	0.088	0.263	0.195
0.092	2.713	0.207	0.028	-0.082	-0.020	0.047	0.069	-0.108	0.148	0.120
-0.001	0.207	2.714	-0.022	-0.098	0.082	-0.075	0.077	-0.092	-0.125	0.053
0.087	0.028	-0.022	3.160	0.255	0.008	-0.017	0.025	0.130	-0.054	-0.100
0.038	-0.082	-0.098	0.255	3.031	-0.157	0.005	0.187	0.173	-0.056	-0.112
0.310	-0.020	0.082	0.008	-0.157	1.318	-0.056	0.019	0.010	0.153	-0.019
0.207	0.047	-0.075	-0.017	0.005	-0.056	**0.574**	0.004	0.004	0.031	-0.004
0.143	0.069	0.077	0.025	0.187	0.019	0.004	3.636	0.137	-0.133	0.058
0.088	-0.108	-0.092	0.130	0.173	0.010	0.004	0.137	2.454	0.070	-0.095
0.263	0.148	-0.125	-0.054	-0.056	0.153	0.031	-0.133	0.070	2.089	0.146
0.195	0.120	0.053	-0.100	-0.112	-0.019	-0.004	0.058	-0.095	0.146	1.437

Selon les données synthétiques utilisées ici, on estime que le taux d'usage du tabac pour le Canada est de 24,241 % avec une erreur-type de racine carrée (0,218)=0,4669 %, tandis que le taux d'usage du tabac pour l'Ontario est de 21,899 % avec une erreur-type de racine carrée (0,574)=0,7576 %. La covariance entre les deux estimations des taux est donnée par 0,207 $(\%)^2$.

Par conséquent, le test de l'égalité des taux donne la statistique suivante :

$$T = (\hat{p}_{ONT} - \hat{p}) \Big/ \sqrt{\hat{\text{var}}(\hat{p}_{ONT} - \hat{p})} = \frac{21.899 - 24.241}{\sqrt{0.574 + 0.218 - 2*(0.207)}} = \frac{-2.342}{0.6148} = -3.81.$$

Veuillez noter que la différence, l'erreur-type et la statistique T sont les mêmes que celles obtenues avec WesVar à l'exemple 1. Nous concluons de nouveau que les totaux sont significativement différents au niveau de confiance de 5 %, puisque -3,81 est inférieur à -1,96 (et que nous partons du principe que T suit une loi de distribution normale).

Il convient de souligner qu'un bug dans SUDAAN 9.0.0 empêche actuellement l'utilisateur de créer le fichier de sortie contenant la matrice de variance-covariance. L'utilisateur doit donner une commande d'impression de la matrice de variance-covariance à l'écran, puis utiliser papier et crayon pour calculer la statistique, ou encore recourir à SUDAAN 8.0.2 ou 9.0.1 pour créer le fichier de sortie, ou encore appliquer la solution n° 3.

VI. Exemple 3 – Utilisation de SUDAAN 9.0.0 ou de SUDAAN 9.0.1 pour illustrer la solution n° 3 :

Malheureusement, la sortie de SUDAAN montrée à l'exemple 2 est assez difficile à lire, ce qui mène à des erreurs éventuelles ainsi qu'à des difficultés d'automatisation du processus. En outre, il est impossible actuellement, à cause d'un bug, de produire le fichier requis dans la version 9.0.0. Au lieu d'avoir recours à la méthode papier et crayon, il est plus facile de modifier le code ci-dessus et d'appliquer la solution n° 3.

```
data smktest;
   set nphsdum3(keep=prc8_cur smc8_2 dhc8_sex wt68 bsw1-bsw500);
   if prc8_cur=. then ontario=.;
   else if prc8_cur=35 then ontario=1;
   else ontario=2;
run;

proc descript data=smktest design=brr;
   weight wt68;
   repwgt bsw1-bsw500;
   class ontario;
   var smc8_2;
   catlevel 1;
   diffvar ontario=(1 2) / name="Ontario vs Rest of Canada: Smk_Rate";
run;
```

Dans l'étape des données indiquée ci-dessus, une variable appelée « ontario » est créée, qui prend la valeur de 1 pour les résidents de l'Ontario et la valeur de 2 pour les non-résidents de l'Ontario. On lance alors PROC DESCRIPT comme dans l'exemple 2, mais la variable de la province est retirée de l'énoncé *tables* et insérée dans un énoncé *diffvar*. On peut utiliser l'énoncé *diffvar* pour spécifier des contrastes linéaires qui sont de simples différences entre les deux niveaux d'une variable de catégorie. Dans cet exemple, la différence entre le taux d'usage du tabac pour l'Ontario (ontario=1) et celui pour le reste du Canada (ontario=2) est le contraste d'intérêt. Voici une petite partie des sorties, montrant les résultats des comparaisons (voir tableau 2).

Tableau 2

One		Contrast Ontario vs Rest of Canada: Smk_Rate
Total	Sample Size	15217
	Weighted Size	24859390.94
	Cntrst Total	-1888634.79
	Cntrst Pct	-3.78
	SE Cntrst Pct	0.99
	T-Test	
	Cont.Pct=0	-3.81
	P-value T-Test	
	Cont.Pct=0	0.0002

Il convient de souligner que la valeur de la différence (Cntrst Pct) diffère de celle présentée aux exemples 1 et 2, de même que la valeur de l'erreur-type présentée. Ce n'est pas une erreur! Cela tient à ce que nous utilisons la solution n^{o} 3, dans laquelle nous comparons l'Ontario et le reste du Canada plutôt que l'Ontario et l'ensemble du Canada.

Néanmoins, le test T présenté (c'est le nom que nous avons donné plus haut à la statistique T_2) et la valeur p qui y est associée sont presque identiques à ceux obtenus dans les exemples 1 et 2. La conclusion demeure la même. De nouveau, il faut se rappeler que ces conclusions sont fondées sur des données synthétiques et qu'il ne faut donc absolument pas s'y fier.

VII. Exemple 4 – Utilisation de Bootvar 3.1 pour illustrer la solution n^{o} 1 :

Le programme Bootvar est un ensemble de macros développées en SAS ou SPSS par les méthodologistes à Statistique Canada pour faciliter le calcul des estimations de la variance en utilisant les poids bootstrap. Les fichiers de microdonnées de l'ENSP étaient accompagnés auparavant des versions antérieures de Bootvar. La version plus générique 3.1 peut maintenant être utilisée pour de nombreuses autres enquêtes de Statistique Canada. Bootvar 3.1 est capable de calculer les variances de totaux, de rapports, de différences entre rapports, de percentiles, de tests du chi carré et de paramètres de régressions linéaires ou logistiques. Examinons l'utilisation de la version SAS du programme.

L'estimation de la variance est exécutée en ***deux étapes*** et entraîne l'utilisation de trois programmes SAS. La ***première étape*** consiste à créer un fichier de données qui contient les variables requises aux fins de l'analyse (premier programme). La ***deuxième étape*** consiste à utiliser BOOTVARE_V31.SAS (et MACROE_V31.SAS) pour estimer les variances.

Durant la première étape, il faut créer les variables dérivées des variables d'entrée. C'est-à-dire qu'il faut créer les variables dichotomiques (souvent appelées variables binaires,

bidons ou 0-1) identifiant les enregistrements qui ont une caractéristique d'intérêt comme le fait d'être un fumeur régulier ou un résident de l'Ontario.

Le fichier analytique doit contenir :

- Les variables nécessaires aux fins de l'analyse (variables dérivées y compris les variables dichotomiques et les variables d'entrée qui n'ont pas besoin d'être modifiées).
- La ou les variable(s) de l'identificateur unique des répondants.
- Au besoin, la ou les variable(s) de ventilation identifiant les groupes pour lesquels on souhaite procéder à une analyse distincte.

Pour calculer le taux d'usage du tabac au Canada et le taux d'usage du tabac en Ontario, il faut créer quatre nouvelles variables dichotomiques. C'est ici que l'utilisateur doit prendre des précautions pour éviter d'obtenir des résultats invalides. Les personnes pour lesquelles la variable SMC8_2 est manquante ne devraient pas contribuer aux estimations du taux d'usage du tabac. C'est dire qu'elles ne devraient contribuer ni à l'un ni à l'autre des deux totaux estimés qui composent chaque taux, soit le nombre de fumeurs réguliers dans chaque domaine et le nombre total de personnes dans le champ d'observation dans chaque domaine. Comme Bootvar calcule ces deux totaux séparément avant d'en calculer l'estimation par quotient, une valeur manquante dans le cas d'une observation pour l'une des variables ne garantit pas que l'observation ne sera pas utilisée pour calculer le total de la deuxième variable. Par conséquent, pour indiquer si une personne est ou n'est pas un répondant valide du Canada et (ou) de l'Ontario, les variables dichotomiques créées à cette fin ne doivent viser que les répondants auxquels on a posé la question SMC8_2 et qui ont donné une réponse valide. La création de deux autres indicateurs dichotomiques permet d'identifier ensuite les fumeurs réguliers parmi ces répondants. Voici le code SAS utilisé pour créer le fichier d'analyse qui sera utilisé dans Bootvar:

```
data in1.nphs_dummy_cyc3;

      %let datafid= "H:\SSMD-DMES\CRAD-
      DARC\Course0438\NPHS_c3_dummy_files\Data\Dumyh356.txt";
      %include "H:\SSMD-DMES\CRAD-
      DARC\Course0438\NPHS_c3_dummy_files\Layout\h356_I.sas";
      /* L'énoncé suivant a été ajouté au fichier h356_i.sas pour
      changer les codes de non-réponse en une valeur manquante:
      if smc8_2 in (6,7,8,9) then smc8_2=.;*/

      if smc8_2=. then canada=.;
      else canada=1;
      if smc8_2=. then ontario=.;
      else if prc8_cur=35 then ontario=1;
      else ontario=0;

      if smc8_2=. then smoker=.;
      else if smc8_2=1 then smoker=1;
      else smoker=0;

      ont_smoker=ontario*smoker;

      keep canada ontario smoker ont_smoker realukey personid wt68;
run;
```

Lorsque le fichier d'analyse est prêt, il faut utiliser le deuxième programme BOOTVARE_V31.SAS :

- pour charger les poids bootstrap

```
data bootwt;

        %let datafid="H:\SSMD-DMES\CRAD-
        DARC\Course0438\NPHS_c3_dummy_files\Bootstrp\bd5h356.txt";
        %include "H:\SSMD-DMES\CRAD-
        DARC\Course0438\NPHS_c3_dummy_files\Bootstrp\Layout\b356_i.sas";
run;
%let bsamp=bootwt;
```

- pour préciser que l'analyse doit être faite au niveau global (sans variable de ventilation)

```
%let classes = .;
```

- pour spécifier les paramètres de l'enquête

```
%let ident = realukey personid;
%let fwgt  = fwgt;
%let bsw   = bsw;
%let R     = 1;
%let B     = 500;
```

- pour préciser les statistiques d'intérêt

- Chaque taux d'usage du tabac est en fait un rapport du nombre estimatif de personnes qui sont des fumeurs réguliers au nombre estimatif total de personnes. Pour comparer le taux pour l'Ontario et celui pour le Canada, nous devons calculer la différence entre les deux rapports. Et c'est pourquoi la façon d'obtenir ce type d'analyse au moyen de Bootvar consiste à utiliser les macros suivantes :

```
%ratio(smoker, canada);
%ratio(ont_smoker, ontario);

%diff_rat(ont_smoker,ontario,smoker,canada);
```

  Il convient de souligner que les deux commandes %ratio ne sont pas requises, à moins que l'utilisateur ne veuille également voir quelles sont les proportions individuelles.

Voici la sortie de la macro %diff_rat :

Variance Estimation for a DIFFERENCE BETWEEN RATIOS
using 500 bootstrap replicates

Num1	Den1	Num2	Den2	Num1 size	Num2 size	Difference of ratios	z	p value	Std. err.	C.V.	Lower limit confidence interval 95%	Upper limit confidence interval 95%
ont_smok	ontario	smoker	canada	887	3666	-0.0234	-3.81	0.0001	0.0061	26.25	-0.0355	-0.0114

Dans cette sortie, la valeur z est la valeur de la statistique T. La valeur p associée à la statistique T est presque identique à celle obtenue dans WesVar dans l'exemple 1. La faible différence tient au fait que WesVar calcule les valeurs p sous l'hypothèse que la statistique T suit la loi de *t* et BootVar, sous l'hypothèse d'une loi normale. (Pour plus de détails sur l'utilisation de ces deux distributions différentes, voir l'annexe 2.) La conclusion demeure la même. Nous devrions conclure, au niveau de confiance de 5 %, à une différence significative entre le taux d'usage du tabac pour le Canada et le taux pour l'Ontario. De nouveau, n'oubliez pas que ces conclusions sont fondées sur des données synthétiques et qu'il ne faut s'y fier d'aucune façon.

VIII. Conclusion

Souvent, les gens souhaitent faire des inférences sur une différence entre une sous-population et l'ensemble de la population en ce qui a trait au taux de fréquence d'une caractéristique donnée. Les personnes qui ont accès aux microdonnées confidentielles de certaines des grandes enquêtes analytiques de Statistique Canada ont besoin de méthodes leur permettant de faire pareilles inférences au moyen des outils logiciels à leur disposition. Ces outils logiciels doivent permettre de calculer l'estimation de la variance en utilisant la méthode du bootstrap puisque les poids bootstrap sont la forme en laquelle des renseignements sur le plan de sondage sont fournis dans le cas de bon nombre des enquêtes de Statistique Canada. Dans le présent document, nous utilisons un ensemble de données synthétiques de l'ENSP pour illustrer trois façons dont on peut procéder pour tirer les inférences au moyen des outils logiciels SUDAAN, WesVar et Bootvar. Aucun de ces progiciels ne peut facilement mettre en application toutes trois solutions, mais au moins l'une d'elles est simple à appliquer pour chacun des progiciels. Nous recommandons plus particulièrement la solution n^{o} 1 pour WesVar et Bootvar (voir les exemples 1 et 4) et la solution n^{o} 3 pour SUDAAN (voir l'exemple 3).

Bibliographie

Phillips, Owen. 2004. " Comment utiliser les poids bootstrap avec Wes Var et SUDAAN." Le Bulletin technique et d'information des Centres de données de recherche. (Automne) 1(2):6-16. Statistics Canada Catalogue no. 12-002-XIF.

Research Triangle Institute (2004). SUDAAN Language Manual, Release 9.0 Research Triangle Park, NC: Research Triangle Institute..

Westat (2002). WesVar 4.2 User's Guide, Rockville, MD.

Annexe 1

Démonstration de l'équivalence de deux hypothèses

La présente annexe montre pourquoi comparer un taux simple dans une sous-population au taux simple dans l'ensemble de la population équivaut à comparer le taux dans la sous-population au taux dans le reste de la population.

Supposons qu'une population est subdivisée en une sous-population A et le reste de la population, représenté par A^c. Soient N_A et N_{A^c} le nombre de personnes dans A et A^c respectivement (de sorte que le nombre de personnes dans l'ensemble de la population est $N = N_A + N_{A^c}$). En outre, soient x_A et x_{A^c} le nombre de personnes présentant la caractéristique à l'étude dans A et A^c respectivement (de sorte que le nombre de personnes présentant la caractéristique dans l'ensemble de la population est $x = x_A + x_{A^c}$). Alors les taux de fréquence de la caractéristique dans la sous-population A, dans la sous-population A^c et dans l'ensemble de la population sont, respectivement, $p_A = {x_A}/{N_A}$, $p_{A^c} = {x_{A^c}}/{N_{A^c}}$ et $p = {x}/{N}$.

Par conséquent, nous avons :

$$
\begin{aligned}
p_A = p &\Leftrightarrow \frac{x_A}{N_A} = \frac{x}{N} \\
&\Leftrightarrow \frac{x_A}{N_A} = \frac{x_A + x_{A^c}}{N_A + N_{A^c}} \\
&\Leftrightarrow x_A(N_A + N_{A^c}) = N_A(x_A + x_{A^c}) \\
&\Leftrightarrow x_A N_{A^c} = N_A x_{A^c} \\
&\Leftrightarrow \frac{x_A}{N_A} = \frac{x_{A^c}}{N_{A^c}} \\
&\Leftrightarrow p_A = p_{A^c}
\end{aligned}
$$

Annexe 2

Loi normale ou loi de *t* pour les variables à tester – le problème ddl

On se rappellera qu'à la section intitulée « Un peu de théorie », nous avons déclaré que la valeur de la statistique à tester peut être comparée aux quantiles appropriés d'une loi normale ou d'une loi de *t*. Cela tient au peu de différences entre la loi normale et la loi de *t* auxquelles la variable à tester est comparée dans le cas des tailles d'échantillon sur lesquelles portent généralement les analyses des enquêtes. Toutefois, si l'on utilise une loi de *t*, ce qui est le cas pour SUDAAN et WesVar, il faut préciser le nombre de degrés de liberté (ddl) de la loi. Pour calculer le nombre de degrés de liberté, il est recommandé habituellement de soustraire du nombre d'unités primaires d'échantillonnage (UPE) contenant des personnes échantillonnées dans la (sous)-population étudiée le nombre de strates contenant les personnes échantillonnées dans la (sous)-population étudiée. Pour les enquêtes de Statistique Canada où les poids bootstrap sont fournis, l'analyste ne dispose pas de renseignements facilement accessibles sur le nombre d'UPE ou de strates. Ainsi, on a tendance à utiliser le nombre « par défaut » de degrés de liberté dans le progiciel, qui est le nombre de poids de rééchantillonnage. Si l'analyse porte sur la plupart des membres de l'échantillon d'enquête, ce nombre « par défaut » sera vraisemblablement une estimation conservatrice, mais si l'analyse porte sur une petite sous-population, où la loi de *t* est probablement une meilleure approximation que la loi normale, ce nombre « par défaut » pourrait être beaucoup trop élevé. Il est toujours utile, lorsqu'on utilise une petite sous-population, de voir si les résultats d'un test seraient différents si le nombre de degrés de liberté d'une variable à tester était réduit; dans l'affirmative, il pourrait être utile de tâcher de trouver une meilleure estimation des ddl que l'estimation « par défaut ».

Nota : Dans l'analyse effectuée dans le présent document, si l'on utilise la solution n^{o} 1 ou la solution n^{o} 2, où un taux dans une sous-population est comparé au taux dans la population complète, le nombre recommandé de degrés de liberté pour la variable à tester sous l'hypothèse d'une loi de t serait le nombre d'unités primaires d'échantillonnage contenant les personnes échantillonnées dans la sous-population moins le nombre de strates contenant les personnes échantillonnées dans la sous-population. Si l'on utilise la solution n^{o} 3, où le taux pour une sous-population est comparé à celui pour une autre sous-population, le nombre d'unités primaires d'échantillonnage contenant les personnes échantillonnées dans la sous-population moins le nombre de strates contenant les personnes échantillonnées dans la sous-population est calculé pour chaque sous-population et la moindre des deux valeurs est le nombre recommandé de degré de liberté. Puisque, comme il est indiqué ci-dessus, l'analyste ne dispose pas de renseignements facilement accessibles sur les UPE et les strates lorsqu'il utilise les poids de rééchantillonnage bootstrap, il doit faire preuve de circonspection en interprétant les résultats de ses inférences lorsqu'il examine une sous-population dont l'échantillon est petit.

Utilisation de poids Bootstrap moyens dans Stata : Une révision de BSWREG

Par James Chowhan et Neil J. Buckley

Résumé

Cet article décrit les modifications apportées à un fichier « bswreg » .ado de Stata qui permet d'estimer la variance à l'aide de poids bootstrap. Cette révision comprend l'ajout de nouvelles fonctions de sortie et d'analyse. La principale fonction ajoutée au programme permet aux chercheurs d'utiliser des poids bootstrap moyens en tenant compte du nombre de poids bootstrap utilisés pour générer ces poids moyens. L'utilité de ce programme est illustrée au moyen de l'ensemble de données de l'Enquête sur le milieu de travail et les employés. La version révisée de la commande « bswreg » demeure un outil souple et convivial, compatible avec une gamme variée de méthodes d'analyse de régression et d'ensemble de données. La commande bswreg et les poids bootstrap fondés sur le plan de sondage ne devraient être utilisés pour l'inférence que si celle-ci est théoriquement valide.

Introduction

Cet article décrit les modifications apportées à la commande « bswreg ». BSWREG est un fichier .ado de Stata qui a été créé pour calculer des estimations de la variance à l'aide de poids bootstrap. Piérard et coll. [2004] ont développé ce programme afin de mettre à la disposition des chercheurs qui se servent de Stata un outil convivial et souple qui peut être utilisé avec les poids bootstrap pour tirer parti de l'information sur le plan de sondage d'enquêtes complexes et de calculer des estimations de la variance d'échantillonnage qui tiennent compte du plan de sondage. Le lecteur consultera Piérard et coll. [2004] pour obtenir des renseignements plus détaillés sur la façon d'utiliser le programme bswreg, ses options uniques et les tests de validation de sa robustesse. Le présent article est rédigé en supposant que le lecteur a pris connaissance du contenu de ce rapport antérieur.

La version révisée du programme ajoute de nouvelles fonctions aux sorties affichées par le programme après l'exécution de la commande, mais, avant tout et par dessus tout, elle permet aux chercheurs d'utiliser des poids bootstrap moyens en tenant compte du nombre de poids utilisés pour les générer. Donc, le programme est conçu pour tenir compte du fait que certaines enquêtes de Statistique Canada fournissent des poids bootstrap moyens. Le programme BSWREG figure à l'annexe 1.

Les données de l'Enquête sur le milieu de travail et les employés (EMTE) sont utilisées pour montrer qu'il est important de tenir compte du fait qu'on utilise un poids bootstrap moyen dans le calcul de l'estimation de la variance fondée sur le plan de sondage, comparativement à la méthode d'estimation utilisée pour les poids bootstrap standard.

II. Brève comparaison des méthodes de pondération bootstrap standard et moyenne

Un grand nombre d'enquêtes de Statistique Canada fournissent un poids final (ou poids de sondage final) et des poids bootstrap que les chercheurs peuvent utiliser pour produire des estimations convergentes des paramètres de population et des variances d'échantillonnage qui tiennent compte du plan de sondage, respectivement.

L'estimateur bootstrap standard de la variance de $\hat{\theta}$, utilisé dans ce programme, est donné par Yeo et coll [1999; 3] :

$$v_B\left(\hat{\theta}\right)=\frac{1}{B}\sum_b\left(\hat{\theta}^*_{(b)}-\hat{\theta}^*_{(.)}\right)^2 \qquad \text{où } \hat{\theta}^*_{(.)}=\left(1/B\right)\sum_b \hat{\theta}^*_{(b)} \ . \qquad (1)$$

Cependant, cette estimateur ne convient pas si les poids utilisés sont les poids *bootstrap moyens*, obtenus en calculant la moyenne de poids bootstrap sur C itérations, habituellement pour assurer la confidentialité des renseignements fournis par les participants à l'enquête.

On pourrait soutenir qu'il faut utiliser les coefficients estimés au moyen du poids final d'échantillon complet représentés par $\hat{\theta}^*_{(.)}$, plutôt que la moyenne des $\hat{\theta}^*_{(b)}$, qui sont les estimations des coefficients générées par l'estimation répétée de $\hat{\theta}$ en utilisant B poids bootstrap. La commande bswreg utilise cette dernière estimation.

En général, on génère les poids bootstrap par tirage aléatoire d'échantillons dans chaque strate d'unités primaires d'échantillonnage, avec remise, de sorte que chaque échantillon sélectionné soit de taille égale au nombre d'unités dans l'ensemble de données; puis, on attribue le poids à chaque unité comprise dans l'unité primaire d'échantillonnage sélectionnée en utilisant la même mise en grappe et le même échantillonnage à plusieurs degrés que ceux utilisés pour générer le poids final (de sondage), et on l'ajuste afin qu'il reflète la probabilité de sélection dans l'échantillon aléatoire. En outre, on donne aux observations ou unités d'échantillonnage sélectionnées dans l'échantillon aléatoire un poids bootstrap positif et aux unités non sélectionnées, un poids nul [Satin et Shastry, 1993]. On répète cet échantillonnage de nombreuses fois afin de générer un ensemble de poids bootstrap suffisamment grand pour qu'il soit convergent; le nombre de fois que le processus est répété est égal au nombre d'échantillons bootstrap. Dans l'équation 1 qui précède, il y a B échantillon bootstrap. Par exemple, dans le cas de l'Enquête nationale sur la santé de la population, il y a B=500 échantillons bootstrap.

Dans le cas de nombreuses enquêtes, cet ensemble final de poids bootstrap (B échantillons) est fourni pour l'analyse de la variance. Cependant, après avoir calculé les poids bootstrap des échantillons, certains programmes d'enquête vont une étape plus loin et calculent la moyenne des poids sur C échantillons bootstrap. En modifiant l'estimateur de la variance présenté à l'équation 1, on obtient l'estimateur bootstrap moyen de la variance suivant:

$$v_{\bar{B}}\left(\hat{\theta}\right)=\frac{C}{B}\sum_b\left(\hat{\theta}^*_{(b)}-\hat{\theta}^*_{(.)}\right)^2 \qquad \text{où } \hat{\theta}^*_{(.)}=\left(1/B\right)\sum_b \hat{\theta}^*_{(b)} \qquad (2)$$

où chaque b^{e} poids bootstrap moyen est égale à la moyenne de C poids bootstrap. Dans cette spécification, le terme $\hat{\theta}^{*}_{(b)}$ s'obtient en utilisant la b^{e} variable de poids bootstrap moyen plutôt que la variable de poids bootstrap standard utilisée dans l'équation 1 [Phillips, 2004 et Yeo et coll., 1999].

Dans le cas du poids bootstrap standard, les répliques bootstrap individuelles, dont certaines peuvent donner un poids nul, ne posent aucun risque de divulgation de renseignements confidentiels. Cependant, si B est grand, l'ensemble des répliques bootstrap standard pourraient être examinées afin de dégager le profil des poids nuls et, par conséquent, d'identifier les observations ou enregistrements appartenant à une grappe particulière. Le poids bootstrap moyen, avec moyennes non nulles, est issu de la pratique consistant à s'assurer qu'au moins un poids ne soit pas nul dans C [Yeo et coll., 1999]. Comme le calcul de poids bootstrap moyens de cette façon permet de masquer l'appartenance à une grappe, il réduit le risque de divulgation de renseignements confidentiels.

Par exemple, dans le cas des données de l'EMTE, le nombre initial d'échantillons pour le calcul des poids bootstrap standard est égal à B=5000. Cependant, pour des raisons de confidentialité, on a décidé de calculer des poids bootstrap moyens sur des groupes de C=50 échantillons. Donc, chacun des 100 poids bootstrap moyens fournis pour l'EMTE est égal à la moyenne de 50 poids bootstrap standard.

En intégrant le nombre entier C dans le numérateur de l'estimateur de la variance, on procède à un ajustement qui réintroduit la variabilité qui avait été éliminée en utilisant un poids bootstrap moyen. Donc, C reflète le fait que l'ensemble de poids bootstrap est un ensemble de poids bootstrap moyens qui ont été calculés sur C itérations [Statistique Canada, 2003]. En outre, l'inclusion du scalaire C dans la révision de BSWREG accroît la portée et la fonctionnalité du programme. L'estimateur de la variance et le programme révisés peuvent être utilisés pour tenir compte de variantes de la méthode des répliques répétées et équilibrées (BRR) standard. Plus précisément, on peut les utiliser pour des enquêtes où l'on ne sélectionne que deux unités primaires d'échantillonnage par strate (dans l'exemple tiré du PISA qui suit, les deux UPE par strate sont des écoles).

Les chercheurs qui souhaitent utiliser des données sur les compétences provenant du Programme international pour le suivi des acquis des élèves (PISA) devraient aussi tenir compte de la variance d'échantillonnage supplémentaire due à l'erreur de mesure inhérente à l'utilisation d'échelles de compétences basées sur des valeurs plausibles pour arriver à un estimateur final de la variance d'échantillonnage (totale). Le programme bswreg n'est utile que si l'on ne se sert pas de données sur les compétences. Consulter Lauzon [2004] pour une discussion de l'estimation de la variance en cas d'utilisation de données sur les compétences fondées sur des valeurs plausibles, telles celles de l'EJET/PISA, comme variables dépendantes. Lauzon examine en détail les situations où il est préférable d'utiliser la méthodes bootstrap plutôt que la méthode des répliques répétées équilibrées (BRR) pour traiter les données du PISA et fournit un programme Stata pour ce genre d'applications.

Le Programme international pour le suivi des acquis des élèves (PISA) et les répliques de Fay, qui peuvent être utilisées pour calculer des estimations sans biais de l'erreur-type pour

accompagner les estimations de population, en sont un exemple. Dans la méthode des répliques répétées équilibrées de Fay, T demi-échantillons sont tirés aléatoirement avec remise, de façon semblable à la procédure décrite plus haut, à partir de chaque strate d'unités primaires d'échantillonnage, de sorte que la taille des échantillons sélectionnés soit égale à la moitié du nombre d'unités dans l'ensemble de données. Puis, les poids finals sont ajustés en multipliant la moitié sélectionnée par (2-K) et l'autre moitié, par K, où K est un nombre compris entre 0 et 1. Dans le cas des données du PISA, K est égal à 0,5 [OCDE, 2001]. L'estimateur de la variance de Fay a la forme qui suit :

$$v_{Fay}\left(\hat{\theta}\right)=\frac{1}{T(1-K)^2}\sum_t\left(\hat{\theta}^*_{(t)}-\hat{\theta}^*_{(.)}\right)^2 \qquad \text{où } \hat{\theta}^*_{(.)}=\left(1/T\right)\sum_t\hat{\theta}^*_{(t)}. \qquad (3)$$

Donc, comme le discute Phillips [2004], un même estimateur de la variance peut être utilisé pour le bootstrap moyen et pour la méthode de Fay. Par exemple, dans l'équation 2, on pourrait fixer C à $C=(1-K)^{-2}$ pour que la méthode de Fay puisse être utilisée. Si l'on utilise l'exemple du PISA, dans l'équation 2, C est égal à 4. Pour une discussion plus approfondie, consulter Phillips [2004] et OCDE [2001]. Les chercheurs qui souhaitent appliquer la méthode de Fay aux données du PISA doivent le faire prudemment, à cause de l'erreur de mesure inhérente aux données sur les compétences basées sur des valeurs plausibles, tel que discuté plus haut.

III. Fonctions révisées

Le programme BSWREG Stata ado révisé offre de nombreuses fonctions supplémentaires utiles (voir l'annexe 2 pour la liste complète), dont la possibilité de tenir compte de l'utilisation de poids bootstrap moyen ou d'autres types de méthodes non standard de répliques répétées équilibrées grâce à l'option c*meanbs*. Cette nouvelle fonction permet de spécifier le nombre d'échantillons bootstrap utilisés pour calculer un poids bootstrap moyen. Dans le cas de l'EMTE, la moyenne des poids bootstrap individuels a été calculée sur des groupes de C=50 échantillons, et, dans ces conditions, l'option c*meanbs* devrait être fixée à 50 (voir l'exemple plus loin).

L'algorithme de dénombrement des itérations bootstrap a été modifié de façon à avertir l'utilisateur de l'achèvement des cinq premières répliques pour qu'il puisse vérifier si les itérations avancent progressivement et ne sont pas « gelées ». Le nouveau dénombrement permettra aussi aux chercheurs de mieux estimer le temps d'achèvement. En outre, l'affichage a été modifié afin d'obtenir une présentation/disposition fixe d'affichage des statistiques.

Plusieurs nouveaux résultats ont également été créés dans e( ) pour la commande d'estimation (e-class) bswreg de Stata, à savoir la variable e(numofbs) qui est disponible après avoir exécuté bswreg et qui contient le nombre d'itérations bootstrap réussies, la variable e(N) qui contient le nombre d'observations dans la régression ordinaire non bootstrapée et la variable e(cmd) qui contient « bswreg ». Tous ces résultats s'ajoutent aux matrices des coefficients estimés e(b) et des variances-covariances bootstrapées e(V) qui continuent d'être disponibles. La commande « ereturn list » doit être utilisée pour afficher d'autres scalaires, macros, matrices et fonctions disponibles lorsqu'on utilise la commande BSWREG.

En plus des nouvelles fonctions susmentionnées, bswreg peut désormais être utilisée avec des commandes de régression supplémentaires, dont, sans s'y limiter, reg, areg, qreg, intreg, ivreg, reg3, probit, biprobit, mlogit, heckprob, heckman, glm ou cox. Le programme fonctionne maintenant avec toutes les commandes de régression qui acceptent les poids. La série de commandes « xt » qui n'acceptent pas les poids ne peut être exécutée avec bswreg. La version révisée du programme bswreg règle aussi les problèmes que pose l'estimation d'équations multiples quand on utilise des étiquettes d'équations Stata complexes et comporte la correction d'une erreur qui survenait quand l'estimation du tout premier poids bootstrap échouait.

IV. Procédure – Un exemple

Le programme Stata révisé est aussi facile à utiliser que le programme bswreg original. Il suffit de copier les fichiers « bswreg.ado » et « bswreg.hlp », qui sont décrit à l'annexe 1, dans le répertoire Stata ADO (taper la commande « adopath » à l'invite de Stata pour afficher une liste des répertoires ado dans lesquels placer ce programme), puis d'utiliser la commande suivante :

bswreg depvar [varlist] weighttype=full_sample_weight [if exp] [in range], cmd(STATA_regression_command) [cmdops(options_for_regression_command)] bsweights(bootstrap_weights_varlist) [cmeanbs(integer)] [level(integer)] [bsci] [saving(path_and_filename[,replace])];

Le soulignement indique les formes abrégées pour appeler les options. Afin d'illustrer l'utilisation de cette commande, servons-nous des données de l'Enquête sur le milieu de travail et l'emploi de 1999 et supposons que l'on veuille étudier l'effet de la taille de l'établissement (petit, moyen ou grand), de la paye par employé, du pourcentage de travailleurs de l'établissement couverts par une convention collective, d'un indicateur faisant la distinction entre les lieux de travail sans but lucratif et ceux exploités en vue d'un bénéfice, et de l'existence dans l'établissement de personnel affecté aux ressources humaines sur la mise en place de régimes de primes au rendement individuel.

Les régimes de primes au rendement individuel est l'un des domaine sur lesquels se concentrent les questions de l'EMTE. La question est : « Votre système de rémunération comprend-il l'un ou l'autre des régimes suivants? [Y compris]…primes au rendement individuel (primes, rémunération à la pièce, commission et option d'achat d'actions) » [Statistique Canada, 2001]. Il s'agit d'une variable binaire dont la valeur est 1 s'il existe des primes au rendement individuel et 0 autrement. L'existence de primes au rendement individuel et les facteurs susceptibles d'influencer leur offre sont les éléments essentiels ici.

Dans notre exemple, la taille de l'établissement est déterminée par le nombre total d'employés à chaque lieu de travail. Les établissements dont l'effectif total varie de 0 à 100 sont considérés comme étant petits, ceux dont l'effectif varie de 101 à 500, comme étant moyens et ceux dont l'effectif est égal ou supérieur à 501, comme étant grands. Il s'agit de la classification utilisée habituellement dans le Système de comptabilité nationale du Canada. En tout, nous définissons trois variables nominales de taille d'établissement. Les petits établissements, qui sont

les plus nombreux, représentent 98,2 % de la population. Viennent ensuite les moyens et les grands établissements, représentant 1,58 % et 0,22 % de la population, respectivement.

La paye par employé représente le rendement moyen, par établissement, de la main-d'œuvre pour le travail et les services du capital humain fournis (payroll_per_person), que l'on calcule en divisant la masse salariale brute par le nombre total d'employés pour chaque établissement.

Le pourcentage de travailleurs d'un établissement couverts par une convention collective est donné par la variable de situation syndicale (pct_union). On suppose que le degré de syndicalisation d'un lieu de travail peut avoir une incidence sur les régimes de primes au rendement individuel offertes.

Le champ d'observation de l'EMTE n'inclut pas le secteur public, mais il comprend les établissements exploités en vue d'un bénéfice et les établissements sans but lucratif du secteur privé (variable binaire nonprft_flag, dont la valeur est 1 s'il s'agit d'un établissement sans but lucratif). En principe, les établissements qui ne visent pas à maximiser leurs bénéfices n'accordent pas la même importance que les autres aux régimes de primes au rendement individuel.

La variable de ressources humaines « hr_in » a pour but de déterminer si une personne est chargée ou non des questions relatives aux ressources humaines au lieu de travail. La question est formulée ainsi : « Lequel des énoncés suivants décrit-il le mieux qui a la responsabilité des questions relatives aux ressources humaines dans cet emplacement? » et les options de réponses sont : « 1) Il y a dans cet établissement un service des ressources humaines distinct formé de plus d'une personne; 2) une personne se consacre à plein temps aux questions relatives aux ressources humaines dans cet établissement; 3) les questions relatives aux ressources humaines font partie des fonctions d'un employé de l'établissement, tel le propriétaire ou l'administrateur; 4) les questions relatives aux ressources humaines relèvent d'une personne ou d'un service dans un autre établissement; 5) les questions relatives aux ressources humaines sont traitées de façon ponctuelle dans cet établissement, c'est-à-dire qu'elles ne sont pas la responsabilité d'une personne en particulier; 6) autre, précisez », où la valeur de la variable hr_in est égale à 1 si le répondant choisit l'option 1, 2 ou 3, et nulle, autrement. Les établissements dotés de personnel interne spécialisé en ressources humaines pourraient être plus susceptibles que les autres de mettre en place des régimes de primes au rendement individuel.

Cet exemple se résume à un problème de régression logistique des primes au rendement individuel sur une série de variables nominales de taille, la paye par employé, la syndicalisation, le souci de la rentabilité et la dotation en personnel des ressources humaines à l'aide des données sur les employeurs de l'EMTE de 1999. Pour commencer, il faut s'assurer d'avoir fusionné correctement le fichier de données analytiques et les fichiers de poids bootstrap appropriés (utilisation de l'identificateur unique approprié). Il n'est pas nécessaire d'attribuer un nom aux poids bootstrap pour utiliser le programme BSWREG. Les 100 poids bootstrap moyens seront tous utilisés dans la régression pour obtenir les erreurs-types fondées sur le plan de sondage. La commande pour l'utilisation de ces poids est la suivante :

bswreg incentives medium large payroll_per_person pct_union nonprft_flag hr_in [pw=wkp_final_wt], cmd(logit) bsweights(wkp_bsw1-wkp_bsw100) cmeanbs(50) level(95); (4)

Les résultats de cette régression sont les suivants :

Output 1

```
. bswreg incentives medium large payroll_per_person pct_union nonprft_flag hr_in
> [pw=wkp_final_wt], cmd(logit) bsweights(wkp_bsw1-wkp_bsw100) cmeanbs(50) level(95) ;


1 bootstraps completed
2 bootstraps completed
3 bootstraps completed
4 bootstraps completed
5 bootstraps completed
25 bootstraps completed
50 bootstraps completed
100 bootstraps completed

Results from BSWREG
-------------------

* The confidence intervals below are based on the normal distribution
```

Var_name	Coef	BSse	BSzstat	BSpvalue	BSilow95	BSiup95
medium	1.038241	0.150151	6.914630	0.000000	0.743950	1.332533
large	1.175536	0.243483	4.828008	0.000001	0.698319	1.652753
payroll_pe	0.000024	0.000004	5.803496	0.000000	0.000016	0.000032
pct_union	-0.930827	0.300417	-3.098455	0.001945	-1.519633	-0.342022
nonprft_fl	-1.089882	0.237791	-4.583371	0.000005	-1.555943	-0.623821
hr_in	-0.283383	0.149053	-1.901218	0.057274	-0.575522	0.008756
_cons	-1.231138	0.168566	-7.303616	0.000000	-1.561520	-0.900755

```
Total bootstraps completed: 100
```

Il s'agit de données de sortie appropriées pour l'inférence, puisque nous avons utilisé les poids bootstrap fondés sur le plan de sondage. Toutes nos variables explicatives ont un effet statistiquement significatif au niveau de confiance de 95 %, sauf la variable de personnel spécialisé dans les ressources humaines (hr_in). À noter la différence entre cette sortie et la sortie bswreg non corrigée pour le poids bootstrap moyen grâce à l'utilisation de cmeanbs(50), c'est-à-dire l'effet sur l'inférence si nous excluons l'option cmeanbs(50) pour les données de l'EMTE quand on utilise les poids bootstrap moyens (voir Output 2).

Output 2

```
. bswreg incentives medium large payroll_per_person pct_union nonprft_flag hr_in
> [pw=wkp_final_wt], cmd(logit) bsw(wkp_bsw*) l(95) ;


1 bootstraps completed
2 bootstraps completed
3 bootstraps completed
4 bootstraps completed
5 bootstraps completed
25 bootstraps completed
50 bootstraps completed
100 bootstraps completed

Results from BSWREG
-------------------
```

```
* The confidence intervals below are based on the normal distribution

  +-------------------------------------------------------------------------------------+
  |   Var_name        Coef       BSse      BSzstat   BSpvalue    BSilow95     BSiup95  |
  |-------------------------------------------------------------------------------------|
  |     medium    1.038241   0.021235    48.893818   0.000000    0.996622    1.079860  |
  |      large    1.175536   0.034434    34.139172   0.000000    1.108047    1.243024  |
  | payroll_pe    0.000024   0.000001    41.036919   0.000000    0.000023    0.000025  |
  |  pct_union   -0.930827   0.042485   -21.909389   0.000000   -1.014097   -0.847558  |
  | nonprft_fl   -1.089882   0.033629   -32.409325   0.000000   -1.155793   -1.023972  |
  |-------------------------------------------------------------------------------------|
  |      hr_in   -0.283383   0.021079   -13.443637   0.000000   -0.324698   -0.242068  |
  |      _cons   -1.231138   0.023839   -51.644360   0.000000   -1.277861   -1.184415  |
  +-------------------------------------------------------------------------------------+

Total bootstraps completed: 100
```

La sortie Output 2 est manifestement problématique. Même si les estimations des coefficients sont les mêmes, comme elles devraient l'être, les erreurs-types sont considérablement plus faibles que dans Output 1. Ce résultat est dû au fait que le facteur scalaire, où C=50, n'est pas inclus dans l'équation 2 et que les variances sont donc sous-estimées d'un facteur C. Autrement dit, les erreurs-types sont sous-estimées d'un facteur $\sqrt{C}$ ou $\sqrt{50}$. Donc, la sortie Outpout 2 mène à des inférences incorrectes, puisque nous sommes portés à conclure que l'existence de personnel affecté aux ressources humaines est également un facteur statistiquement significatif au niveau de confiance de 95 %.

On notera que, dans la commande Output 2 qui précède, la liste des variables de poids bootstrap est spécifiée comme étant « wkp_bsw* ». Les chercheurs pourraient juger utile d'utiliser le joker, ou astérisque, lorsqu'ils spécifient une liste de variables qui ne sont pas nécessairement en ordre numérique dans l'ensemble de données Stata utilisé. Cette option permet d'éviter que l'algorithme intégré dans Stata pose un problème, parce qu'il est conçu pour sélectionner les variables sur la fourchette spécifiée dans l'ordre où elles surviennent dans l'ensemble de données, plutôt que dans l'intervalle logique sous-entendu par les bornes. Par exemple, si les bornes sont « bsw1-bsw100 » et que les quatre premiers (des cent) poids spécifiés dans l'ensemble de données sont bsw1, bsw10, bsw100 et bsw2, si l'énoncé de la liste de variables *varlist* est « bsw1-bsw100 », dans toute commande Stata, seules les trois premières variables (poids) seront sélectionnées (bsw1, bsw10, bsw100), plutôt que la gamme complète. Par contre, si *varlist* est énoncée comme étant « bsw* », la gamme complète des 100 poids sera sélectionnée.

Output 3

```
> logit incentives medium large payroll_per_person pct_union nonprft_flag hr_in
> [pw=wkp_final_wt];


(sum of wgt is   7.1789e+05)

Iteration 0:   log pseudo-likelihood = -3817.6905
Iteration 1:   log pseudo-likelihood = -3635.8102
Iteration 2:   log pseudo-likelihood = -3631.4552
Iteration 3:   log pseudo-likelihood = -3631.4242
Iteration 4:   log pseudo-likelihood = -3631.4242

Logit estimates                                   Number of obs   =       6271
                                                  Wald chi2(6)    =     103.01
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -3631.4242                Pseudo R2       =     0.0488
```

```
------------------------------------------------------------------------------
             |               Robust
  incentives |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      medium |   1.038241   .1580988     6.57   0.000     .7283732    1.348109
       large |   1.175536   .2388543     4.92   0.000     .7073897    1.643681
payroll_pe~n |   .0000242   3.82e-06     6.33   0.000     .0000167    .0000317
   pct_union |  -.9308272   .2927079    -3.18   0.001    -1.504524   -.3571304
nonprft_flag |  -1.089882   .2453014    -4.44   0.000    -1.570664   -.6091007
       hr_in |   -.283383   .1420044    -2.00   0.046    -.5617065   -.0050594
       _cons |  -1.231138   .1660942    -7.41   0.000    -1.556676   -.9055991
------------------------------------------------------------------------------
```

Il importe aussi de souligner que la régression logit avec erreurs-types robustes donnerait également lieu à une inférence incorrecte, parce que les poids bootstrap ne sont pas utilisés du tout. D'après l'information qui figure dans la sortie Output 3, il semble que toutes les variables soient significatives au niveau de confiance de 95 %, mais que les erreurs-types soient biaisés et donnent lieu à une inférence inappropriée. La sortie générée par l'équation 4 (Output 1) contient les erreurs-types fondées sur le plan de sondage et les valeurs p connexes.

Le programme est utile non seulement pour les techniques de régression, mais peu aussi être utilisé pour calculer diverses statistiques sommaires, comme des fréquences, des moyennes et des ratios. Voir Piérard et coll. [2004] pour une discussion des limites et pour des exemples de calcul de ces statistiques.

V. Conclusion

Le programme décrit est axé sur la production d'estimations de la variance fondées sur le plan de sondage et appropriées pour l'inférence dans le cas de diverses enquêtes sociales de Statistique Canada. Il peut désormais être appliqué aux données de toute enquête pour laquelle sont produits des poids bootstrap, c'est-à-dire une grande gamme d'ensembles de données provenant de l'Enquête sociale générale (ESG), de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), de l'Enquête nationale sur la santé de la population (ENSP), de l'Enquête sur la dynamique du travail et du revenu (EDTR), de l'Enquête sur le milieu de travail et l'emploi (EMTE) et, avec certaines contraintes, du Programme international de suivi des acquis des élèves (PISA) et de l'Enquête sur les jeunes en transition (EJET), pour n'en nommer que quelques-uns.

Les modifications apportées au programme visent à étoffer les fonctions existantes et à continuer d'offrir aux chercheurs qui utilisent Stata un outil souple, convivial et exact.

Bibliographie

Lauzon, Darren. 2004. " Estimation de la variance dans le cas de données sur les compétences basées sur des valeurs plausibles : Deux programmes STATA pour l'analyse des données de l'EJET/PISA." Le Bulletin technique et d'information des Centres de données de recherche. (Printemps) 1(1):33-59. Statistique Canada, catalogue no. 12-002-XIF.

Organization for Economic Co-operation and Development (OECD). 2001. "Manual for the PISA 2000 Database." Programme international pour le suivi des acquis des élèves (PISA) 2000.

Phillips, Owen. 2004. " Comment utiliser les poids bootstrap avec Wes Var et SUDAAN." Le Bulletin technique et d'information des Centres de données de recherche. (Automne) 1(2):6-16. Statistique Canada, catalogue no. 12-002-XIF.

Piérard, Emmanuelle, Neil Buckley, and James Chowhan. 2004. " Pour une utilisation plus conviviale de la méthode bootstrap : fichier ADO dans Stata." Le Bulletin technique et d'information des Centres de données de recherche. (Printemps) 1(1):15-32. Statistique Canada, catalogue no. 12-002-XIF.

Satin, Alvin, and Wilma Shastry. (1993) Survey Sampling: A Non-Mathematical Guide. Ottawa: ministère de l'industrie, Statistique Canada, Division des méthodes d'enquêtes sociales. Catalogue No. 12-602-XPE.

Statistique Canada. 2003. " Guide pour l'analyse de L'Enquête sur le milieu de travail et les employés 2001" Division de l'analyse des entreprises et du marché du travail & Division de la statistique du travail. (Juin) Ottawa.

Statistique Canada. L'Enquête sur le milieu de travail et les employés (EMTE), 1999. Division de l'analyse des entreprises et du marché du travail et Division de la statistique du travail.. 4-4700-2.1: 1999-04-01. STC/LAB-075-75055. Ottawa: Statistique Canada.

Yeo, Douglas, Harold Mantel, and Tzen-Ping Liu. 1999. "Bootstrap Variance Estimation For the National Population Health Survey." American Statistical Association: Proceedings of the Survey Research Methods Section. Baltimore, August.

Annexe 1

Fichier ado

```
*                                   WARNING
* The authors are the owners of all intellectual
* property rights (including copyright) in this software.  Subject to the terms below,
* you are granted a non-exclusive and non-transferable license to use this software.
*
* This software is provided "as-is", and the owner makes no warranty, either express
* or implied, including but not limited to, warranties of merchantability and fitness
* for any particular purpose.  In no event will the owner be liable for any indirect,
* special, consequential or other similar damages.  This agreement will terminate
* automatically without notice to you if you fail to comply with any term of this
* agreement.

* TO CHANGE THE DECIMAL DISPLAY FORMAT OF THE BOOTSTRAPPED OUTPUT SEARCH FOR THE "FORMAT" COMMAND NEAR
THE BOTTOM OF THIS PROGRAM;

program define bswreg, eclass sortpreserve byable(recall)

* March 1st, 2005 Buckley, Chowhan
*    Reset variable and equation labels since those with spaces were interfering with some regression
*       commands like ologit etc., if you are using Stata versions prior to 8.2 you will need to drop
spaces from variable labels before running bswreg
* October 21st, 2004 Buckley, Chowhan
*    BSWREG should now work with any regression command that accepts a weight
*       (including, but not limited to commands like: reg, qreg, intreg, ivreg, reg3, probit, biprobit,
heckprob, heckman, glm etc...)
*     fixed problem with running BSWREG with regression methods that analyze censored data containing
missing values (e.g. INTREG is now fully functional within BSWREG)
* September 30th, 2004 Buckley, Chowhan
*    added possibility of mean bootstrap weights
*    changed bootstrap count algorithm
*     created e(numofbs) variable that is available after running bswreg and contains the number of
bootstraps successfully run
*    created e(N) variable that contains number of observations in plain unbootstrapped regression
*    created e(cmd) variable that contains "bswreg"
*    fixed statistic display format/layout
* August 8th, 2003 Pierard, Buckley, Chowhan (original)

# delimit;
version 7.0;


syntax anything [aweight pweight fweight iweight] [if] [in], cmd(string) [cmdops(string)]
  BSWeights(varlist numeric) [Cmeanbs(integer 1)] [Level(integer 95)] [bsci] [SAVing(string)];


*This sets the touse variable = 1 if observation is in our sample;
marksample touse;
*Error check to make sure a weight was used;
if "`weight'"==""
  {;
  noi di in red "BSWREG error: You must specify a weight!";
  exit;
  };

quietly
{;

*Preserve the original dataset and set parameter values and setup temporary matrices;
preserve;
set more 1;
tempvar esamplevar;
tempname bhat bsVC bsbhat bsbetas;

*The next line runs the wanted regression and checks for errors;
capture `cmd' `anything' [`weight'`exp'] `if' `in', `cmdops';

if _rc ~= 0
   {;
   noi di in red " ";
   noi di in red "Error doing: `cmd' `anything' [`weight'`exp'] `if' `in', `cmdops'";
   noi di in red " ";
   noi di in red "The regression command you have typed in resulted in an error, please investigate";
   noi di in red "this error outside of the 'bswreg' program by typing in the regression command
itself";
   noi di in red "with the options you specified.";
   noi di in red " ";
   exit;
   };

*The next line removes all variable and equation labels because they will cause problems if they
contain spaces (they will be put back later);
capture label language bswreg, new;
*The next line runs the wanted regression and we store the coefficients in a matrix for later use;
```

```
`cmd' `anything' [`weight'`exp'] `if' `in', `cmdops';
local _numofobs = e(N);
gen `esamplevar'=e(sample);

*e(b) is a 1x(k+1) coefficient vector if the model has a constant and k is the number of variables
other than the constant;

matrix `bhat'=e(b);
matrix list `bhat';
matrix `bsVC'=e(V);
*The next line initializes the bootstrap coefficients matrix with the original sample weighted
coefficients to get the correct matrix dimensions, this first column will be removed later;
matrix `bsbetas'=(`bhat'');

*we store the variable names of the regressors and the number of regressors in local macros;
local _varnames : colfullnames(`bhat');
local _k=colsof(`bhat')-1;
local _k1=`_k'+1;

*Generate concatenated list of placeholder regressor variable names xc1-xck1, later to be turned into
variables;
local _xclist="";
forvalues _i = 1/`_k1'
   {;
   local _xclist `_xclist' _xc`_i';
   };
*We assigned these placeholder variable names to the regressors in the coefficient vector;
matrix colnames `bhat' = `_xclist';
*Each "true estimate of beta" is saved under it's own variable name;

svmat double `bhat', name(col);
matrix colnames `bhat' = `_varnames';

*Realboot is the actual number of successful bootstrap regressions run in case we get any
convergence/regression errors etc., it starts off at the specified number of bootstrap weights;
local _realboot: word count `bsweights';
noi di " ";

*The main bootstrap loop will run with each bootstrap weight in the supplied bsweight varlist and exit
with the matrix named BETAS containing all the bootstraps of our coefficients, a (boot)x(k+1)
dimensional matrix;
local _i 1;
*Start of bootstrap loop;
foreach bswvar of local bsweights
 {;

 *Display notice of number of completed bootstraps every time 50 are completed;
 if (mod(`_i',100)==0 | `_i'<6 | `_i'==25 | `_i'==50)
  {;
  noi di in green `_i' " bootstraps completed";
  };

 *Run the regression with the chosen set of bootstrap weights, only use the coefficients if there are
no errors;

 capture `cmd' `anything' [`weight'=`bswvar'] `if' `in', `cmdops';
 if _rc==0
  {;

   *Store coefficients in the bootstrap matrix;
   matrix `bsbhat'=get(_b);
   *bsbhat is a 1x(`k'+1) (row) vector if the model has a constant.  Need to transpose;
   matrix `bsbhat'=`bsbhat'';

   *If we have the proper number of coefficients then add them to the bootstrap matrix, otherwise do
not add them (this most likely arises due to a regressor being dropped due to multicollinearity;
   if rowsof(`bsbhat')==`_k1'
        {;
        *Append the coefficients from the current bootstrap to the aggregate matrix;
        matrix `bsbetas'=(`bsbetas',`bsbhat');
        };
   else
        {;
        matrix drop `bsbhat';
        local _realboot=`_realboot'-1;
        noi di "Bootstrap #`_i' has been dropped for not having the correct number of coefficients";
        };
  };
 else
  {;
  local _realboot=`_realboot'-1;
  noi di "bootstrap #`_i' has been dropped due to an error estimating the regression";
  };
 local _i=`_i'+1;
 };
*Now we remove the initial column of coefficients that used the original sample weight;
matrix `bsbetas'=`bsbetas'[1...,2...];
*End of bootstrap loop;

*All the bootstraps have been completed now calculate the new standard errors and display relevant
statistics;
*We must transpose the matrix to make each row now, then column, a new variable;
```

```
matrix `bsbetas'=`bsbetas'';
*Generate concatenated list of colnames, later to be turned into variables;
local _xvlist="";
forvalues _i = 1/`_k1'
   {;
   local _xvlist `_xvlist' _xv`_i';
   };

*Calls each row of the matrix by the name of the independent variable it corresponds to (we call them
_xv'_i' so that they are not mixed up with the "real" variables);
matrix colnames `bsbetas'= `_xvlist';

*Separate each column as a new variable.  The format of the data must be specified.  It renames each
variable by the name of the column;
svmat double `bsbetas', name(col);

*Generate the bootstrapped variance-covariance matrix, you can access this in e(V) after running the
BSWREG ado file;
*CmeanBS is the number of bootstrap weight samples used to calculate an average bootstrap weight
sample;
*When CmeanBS is not equal to 1 a mean bootstrap factor exists dependent on the survey, the default
value is 1;
forvalues _i = 1/`_k1'
 {;
 forvalues _j = 1/`_k1'
  {;
    correlate _xv`_i' _xv`_j', covariance;
    matrix `bsVC'[`_i',`_j'] = (((`_realboot'-1)*(`cmeanbs'))/`_realboot')*r(cov_12);
  };
 };

*Generate the standard deviation, t-stat, conf. int. etc. for each variable;
tempvar _bsobs _uniqobs _coefnum;
gen `_bsobs'=_n;
forvalues _i = 1/`_k1'
 {;

 sum _xv`_i';
 * Like the SAS bootvar program, we use (boot-1)/boot because variance and standard error have
different denominators;

 * See above for description of CmeanBS;
 gen _sdx`_i'=sqrt((((`_realboot'-1)*(`cmeanbs'))/`_realboot')*r(Var)) in 1/1;
 gen _t`_i'=_xc`_i'/_sdx`_i' in 1/1;
 gen _abst`_i'=abs(_t`_i') in 1/1;
 gen _p`_i'=2*norm((-1)*_abst`_i') in 1/1;
* gen _p`_i'=2*ttail(`_realboot'-1,_abst`_i') in 1/1;
 if "`bsci'"==""
 {;
   gen _low`level'`_i'=_xc`_i'-invnorm(1-((1-(`level'/100))/2))*_sdx`_i';
   gen _high`level'`_i'=_xc`_i'+invnorm(1-((1-(`level'/100))/2))*_sdx`_i';
 };
 if "`bsci'"=="bsci"
   {;
   sort _xv`_i';
   local _obslow= max(1,round(((1-(`level'/100))/2)*`_realboot',1));
   local _obshigh= max(1,round((1-((1-(`level'/100))/2))*`_realboot',1));
   local _obslow2= _xv`_i'[`_obslow'];
   local _obshigh2= _xv`_i'[`_obshigh'];
   sort `_bsobs';
   gen _low`level'`_i'= `_obslow2' in 1/1;
   gen _high`level'`_i'= `_obshigh2' in 1/1;
   };
 };

*Assign each coefficient its true regressor name stored at the beginning of this program;
local _i=1;
foreach _curname in `_varnames'
 {;
 gen str10 _xname`_i'="`_curname'";
 local _i=`_i'+1;
 };

*Reshape the data so that the bootstrapped stats can be displayed easily, and then display the
results;
keep _xname* _xc* _sdx* _t* _p* _low`level'* _high`level'*;
drop if _n>1;
gen `_uniqobs'=1;

reshape long _xname _xc _sdx _t _p _low`level' _high`level', i(`_uniqobs') j(`_coefnum');

*The %9.4f tells stata to display the bootstrapped results to 6 decimals using 15 numbers total --
this can be changed to suit tastes;
format _xc _sdx _t _p _low`level' _high`level' %11.6f;
*creates nice labels for variables
label var _xname "Name of variable";
ren _xname Var_name;
label var _xc "Coefficient estimate";
ren _xc Coef;
label var _sdx "Bootstrap standard error of coefficient";
ren _sdx BSse;
label var _t "Bootstrap z-statistic";
```

```
ren _t BSzstat;
label var _p "Bootstrap p-value";
ren _p BSpvalue;
if "`bsci'"==""
 {;
 label var _low`level' "Bootstrap lower confidence interval assuming a normal distribution";
 label var _high`level' "Bootstrap upper confidence interval assuming a normal distribution";
 };
if "`bsci'"=="bsci"
 {;
 label var _low`level' "Bootstrap lower confidence interval using bootstrap sample distribution";
 label var _high`level' "Bootstrap upper confidence interval using bootstrap sample distribution";
 };
ren _low`level' BSilow`level';
ren _high`level' BSiup`level';

*Display RESULTS!;
noi display in green " ";
noi display in green "Results from BSWREG";
noi display in green "-------------------";
noi display in green " ";
if "`bsci'"=="bsci"
 {;
 noi display in green "* The confidence intervals below are based on the bootstrapped distribution";
 };
else noi display in green "* The confidence intervals below are based on the normal distribution";
*noi display in green " ";
format Coef BSse BSzstat BSpvalue BSilow`level' BSiup`level' %10.6f;
format Var_name %10s;
noi list Var_name Coef BSse BSzstat BSpvalue BSilow`level' BSiup`level', nodisplay noobs;

noi di " ";

noi di "Total bootstraps completed: `_realboot'";

*Set the eclass variables like the coefficients and the variance-covariance matrix into their
appropriate matrices so that F-tests and the like can be run;
*If you wish the TEST command to produce F-tests after the BSWREG command then add ",
dof(`_realboot')" to the line below;
estimates post `bhat' `bsVC';
*This next line creates a e(numofbs) scalar available after running bswreg that contains the number of
boostraps run, di e(numofbs);
estimates scalar numofbs = `_realboot';
estimates scalar N = `_numofobs';
estimates local cmd = "bswreg";

*Save the bootstrap raw data is the "SAVING" option has been used;
if "`saving'"~=""
 {;
  drop _*;
  save "`saving'", `replace';
 };

*The next line removes all variable and equation labels because they will cause problems if they
contain spaces (they will be put back later);
capture label language default;

*Restore the original dataset
restore;

};
end;
```

BSWREG help file

```
{smcl}
{* 21October2004 Buckley/Chowhan}

{* 30September2004 Buckley/Chowhan}
{* 8August2003 Pierard/Buckley/Chowhan}
{hline}
help for {hi:BSWREG}
{hline}

{title:BSWREG - uses bootstrap weights to calculate standard errors in models involving complex survey
data.}

{p 8 13}{cmd:bswreg} depvar [varlist] {it:weighttype}={it:full_sample_weight} [{cmd:if} {it:exp}]
[{cmd:in}              {it:range}][{cmd:,}              {cmd:cmd(}{it:STATA_regression_command}{cmd:)}
[{cmd:cmdops(}{it:options_for_regression_command}{cmd:)}]
  {cmdab:bsw:eights(}{it:bootstrap_weights_varlist}{cmd:)}       [{cmdab:c:meanbs(}{it:integer}{cmd:)}]
[{cmdab:l:evel(}{it:integer}{cmd:)}]                                                    [{cmd:bsci}]
[{cmdab:sav:ing(}{it:path_and_filename}[{cmd:,replace}]{cmd:)}];
{p} {cmd:cmd()} and {cmd:bsweights()} are required options for the {cmd:BSWREG} command.
{p} {cmd:by ...: } and {cmd:bysort ...:} can be used with {cmd:BSWREG}. See help {help by}.
{p} {cmd:aweight}s, {cmd:fweight}s, {cmd:iweight}s, and {cmd:pweight}s are allowed as long as the
      given regression command is compatible with them. See help {help weights}.
```

```
{p} As {cmd:BSWREG} is an eclass STATA program, it provides STATA with the {cmd:e(b)} coefficient
vector and the {cmd:e(V)} bootstrapped variance-covariance matrix.
    The {cmd:test} command can be used immediately following the {cmd:BSWREG} command to conduct Wald
tests based on the chi-squared distribution.

{inp:The software is provided "as-is" and the authors are not responsible for any misuse.}

{title:Description}

(used to calculate regression statistics using Statistics Canada's bootstrap weights)

{p}{cmd:bswreg} runs a number of regressions, each with a particular bootstrap
weight so that bootstrapped standard errors on the coefficients can be calculated
and displayed. Use of bootstrap weights is recommended for calculating reliable
standard errors, confidence intervals etc. on data from complex
household surveys.

The user provides the names of the bootstrap weights to the {cmd:BSWREG} command
in the {cmdab:bsw:eights(varlist)} option. You must already have the appropriate
bootstrap weights merged into your datafile for this command file to work.  NPHS
merges on REALUKEY and SLID merges on PERSONID. Below is a sample .DO file that
merges NPHS bootstrap weights into a datafile named data.dta:

{inp:use data.dta, replace"}
{inp:sort realukey"}
{inp:save data.dta, replace"}
{inp:use bootstrap/sas_bs_wt_1_4.dta, replace"}
{inp:destring realukey, replace"}
{inp:sort realukey"}
{inp:merge realukey using data.dta"}
{inp:keep if _merge==3"}

{title:Options}

{p  0  4}{cmd:cmd(}{it:STATA_regression_command}{cmd:)} specifies the Stata regression command to
bootstrap. This is a {cmd:required} option. "regress", "probit" and "logit" are a few possibilities.

{p 0 4}{cmd:bsweights(}{it:varlist}{cmd:)} specifies a variable list of the bootstrap weight names.
This is a {cmd:required} option. For instance, if your bootstrap weights are named bsw1 to bsw500, you
may wish to use the
{cmd:bsweights(bsw1-bsw500)} option. In order to avoid Stata variable ordering problems it might be
better to specify {cmd:bsweights(bsw*)} when using all weights.

{p 0 4}{cmd:cmdops(}{it:options_for_regression_command}{cmd:)} specifies the options you wish to use
on the Stata regression command provided in {cmd:cmd()}. Some options are useful and others are
meaningless in a bootstrap weighting context.
For instance, if you wish to run the REGRESS command with no constant then use the {cmd:cmd(regress)
cmdops(noconstant)} options.  Options like {cmd:robust} are meaningless in this context since the
command computes bootstrap weighted
standard errors not robust ones.

{p 0 4}{cmd:cmeanbs(}{it:integer}{cmd:)} specifies the number of bootstrap weight samples each mean
bootstrap
weight is averaged over, in the case of surveys that use mean boostrap weights. The default is
that the bootstraps privided are not mean bootstrap weights, {cmd:cmeanbs(1)}.

{p 0 4}{cmd:level(}{it:integer}{cmd:)} specifies the confidence level, in percent,
for confidence intervals.  The default is {cmd:level(95)}. See help {help level}.

{p 0 4}{cmd:bsci} specifies that the confidence intervals be calculated from the raw bootstrapped
distribution of coefficients rather than using the standard formula
based on the bootstrapped standard error and the normal distribution.

{p 0 4}{cmd:saving(}{it:filename}[{cmd:,replace}]{cmd:)} saves the bootstrap statistics in a
separate Stata dataset file that can later be loaded and used by other .DO and .ADO files.
If you do not specify an extension, {cmd:.dta} will
be assumed. Include the {cmd:,replace} option to overwrite an existing file.

{title:Outputed variables}
{inp: Var_name:}  This is the STATA variable name of the regressor.
{inp: Coef:}      This is the coefficient from the specified regression.
{inp: BSse:}      This is the new standard error of the coefficient,
            calculated using bootstrap weights.
{inp: BSzstat:}   This is the new z-stat of the coefficient,
            calculated as the coefficient divided by the bootstrapped standard error.
{inp: BSpvalue:}  This is the new p-value of the coefficient,
            calculated using the z-statistic.
{inp: BSilow(level):} This is the lower (level)% confidence interval around the coefficient
            using the bootstrapped std. error.
{inp: BSiup(level):}  This is the upper (level)% confidence interval around the coefficient
            using the bootstrapped std. error.

{title: e-class results}
{inp: e(numofbs): Scalar} The number of successful bootstrap replications.
{inp: e(N): Scalar}       The number of observations in the underlying survey sample.
{inp: e(cmd): Macro}      Contains "bswreg".
{inp: e(b): Matrix}       This is the vector of coefficients.
{inp: e(V): Matrix}       This is the bootstrapped variance-covariance matrix.

{title:Examples}

{p 8 12}{inp:. bswreg income education rural [aw=wt] if married==1, cmd(regress) bsw(bsw1-bsw500)}
```

```
{p 8 12}{inp:. bswreg employed education rural [aw=wt66], cmd(probit) bsw(bsw50-bsw100)}
{p 8 12}{inp:. bysort maritalstatus: bswreg income education rural [aw=wt], cmd(reg) bsw(bsw1-bsw500)}
  {inp:cmdops(noconstant) level(99) bsci saving(c:\data\bsw1.dta,replace)}
{p 8 12}{inp:. bswreg wesemployeeincentives wesworksize [aw=wt], cmd(logit) bsw(bsw1-bsw500)
cmeanbs(50)}
```

Annexe 2

La commande bswreg permet d'utiliser différentes options. Le programme en offre plusieurs :

cmd : précise la commande de régression de Stata pour la méthode bootstrap. La sélection de cette fonction est nécessaire. Les commandes de régression qui suivent ont été mises à l'essai explicitement : regress, logit, probit, tobit, ologit, oprobit, biprobit, mlogit, qreg, glm, intreg, boxcox, (à peu près toute technique d'estimation en une étape devrait fonctionner avec ce programme) et les commandes « xt » ne comportant pas deux étapes pour lesquelles il est possible d'utiliser les poids.

bsweights : établit une liste de variables correspondant aux noms des poids bootstrap. La sélection de cette fonction est nécessaire. Par exemple, si les poids bootstrap sont nommés bsw1 à bsw500, on pourrait utiliser la spécification bsweights(bsw1-bsw500). Pout éviter tout problème dû à l'ordre des variables dans Stata, il est préférable de spécifier bsweights(bsw*) si l'on veut utiliser tous les poids.

cmdops : permet de spécifier, dans la commande de régression Stata fournie dans cmd(), les options que l'on souhaite utiliser. Certaines options sont utiles et d'autres n'ont aucune pertinence dans le contexte d'une pondération bootstrap. Par exemple, pour exécuter la commande REGRESS sans constante, on utilisera les options cmd(regress) cmdops(noconstant). Les options telles que « robust » sont inutiles ici, puisque la commande calcule des erreurs-types pondérées selon la méthode bootstrap et non des erreurs-types robustes.

cmeanbs : sert à préciser le nombre de poids bootstrap utilisés pour calculer un poids bootstrap moyen; les facteurs de pondération bootstrap moyens dépendent de l'enquête. La valeur par défaut de cette option est 1, ce qui signifie que les poids bootstrap ne sont pas des poids bootstrap moyens.

level : permet de préciser le niveau de confiance, en pour cent, pour les intervalles de confiance. La valeur par défaut est level(95).

bsci : indique que les intervalles de confiance doivent être calculés d'après la distribution brute des coefficients obtenus par la méthode bootstrap, plutôt qu'avec la formule stantard fondée sur la loi normale et l'erreur-type calculée par la méthode bootstrap. Cette option s'adresse aux utilisateurs qui pourraient avoir des raisons théoriques d'utiliser les intervalles de confiance calculés d'après la distribution des coefficients estimés par la méthode bootstrap.

saving : sert à sauvegarder les statistiques bootstrap dans un fichier de données Stata distinct qui peut être chargé par la suite et utilisé par d'autres fichiers .DO et .ADO. Si l'extension n'est pas

précisée, celle attribuée par défaut sera .dta. Cette fonction comprend l'option « replace » pour écraser un fichier existant.

Note technique

Le ménage comme unité d'analyse dans l'Enquête longitudinale nationale sur les enfants et les jeunes

Par Franck Larouche et Charles Tardif

L'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) vise à suivre jusqu'à l'âge adulte le développement et le bien-être d'un échantillon représentatif d'enfants canadiens âgés de 0 à 11 ans au premier cycle (1994-1995). Cette enquête répétée tous les deux ans a été conçue de manière à ce que l'enfant soit l'unité d'analyse. Les poids transversaux et longitudinaux attribués à chaque enregistrement sont des poids correspondant à l'unité d'analyse, soit l'enfant.

Les enfants dans l'échantillon de l'ELNEJ sont sélectionnés à partir d'un plan complexe afin de répondre à différents besoins et ce, en tenant compte de certaines contraintes opérationnelles. Une partie de l'échantillon provient de l'Enquête sur la population active (EPA) et une autre, du Registre des naissances. En plus de l'échantillon initial du cycle 1, mentionnons également qu'un nouvel échantillon d'enfants de 0 et 1 an est sélectionné à tous les cycles subséquents. Le poids initial de l'enfant lors de la sélection de l'échantillon correspond, sommairement, à l'inverse de sa probabilité de sélection. Par la suite, ce poids est ajusté pour tenir compte de la non-réponse totale lors de la collecte en utilisant des caractéristiques relatives à l'enfant. Suite à l'ajustement de non-réponse, les poids sont post-stratifiés par province, âge et sexe de l'enfant pour qu'ils représentent les totaux démographiques connus par province, âge et sexe.

Certains éléments d'information ont été recueillis au niveau du ménage, mais il n'est pas possible de faire des estimations généralisables à l'ensemble des ménages canadiens. Toutes les inférences doivent être faites au niveau de l'enfant. Par exemple, nous pouvons estimer le nombre d'enfants vivant dans un ménage avec un seul parent mais nous ne pouvons pas estimer le nombre de ménages avec un seul parent.

Prendre le poids moyen des enfants appartenant à un même ménage comme une variable de poids au niveau du ménage et faire des estimations qui se diraient représentatives des ménages canadiens n'est pas recommandé. En effet, de tels poids ne seraient pas ajustés pour être représentatifs de l'ensemble des ménages. En fait, toute technique de transformation du poids de l'enfant pour en faire un poids ménage ou autre, qui ne tiendrait pas compte des différents ajustements nécessaires pour la non-réponse et la post-stratification, n'est pas souhaitable.

Il est à noter également que la stratégie d'échantillonnage a été modifiée au fil des cycles. Lors du cycle 1, jusqu'à quatre enfants par ménage ont été choisis. Au cycle 2, réalisant le lourd fardeau de réponse des ménages avec plusieurs enfants, il a été décidé de réduire à deux le nombre maximal d'enfants sélectionnés par ménage. Lors du cycle 3, seuls des enfants âgés de 0 an ont été sélectionnés à partir de l'EPA. Donc, mis à part quelques exceptions (principalement

les jumeaux), un seul nouvel enfant a été choisi par ménage. Pour leur part, les enfants de 1 an et 5 ans nouvellement choisis au cycle 3 l'ont été à partir du Registre des naissances et ainsi, un seul enfant par ménage a été choisi. Au cycle 4, les enfants de 0 et 1 an ont été choisis à partir de l'EPA. Il est donc possible que certains ménages aient deux enfants sélectionnés (les jumeaux ou certains cas où il y avait deux enfants dans le ménage qui avaient moins de 2 ans) mais dans la grande majorité des cas, un seul enfant par ménage a été choisi. Aux cycles 5 et 6, la stratégie d'échantillonnage a été modifiée afin de ne sélectionner qu'un seul enfant par ménage. Dans le cas de jumeaux, un seul des deux enfants a été choisi par l'ELNEJ. La présence de plus d'un enfant par ménage est donc de moins en moins importante.

Dans le cas échéant, quoiqu'il semble étrange d'inclure deux enregistrements d'un même ménage, il n'y a aucun problème à le faire. Les poids ont été obtenus pour produire des estimations de la population d'enfants canadiens et la variance calculée avec les poids de réplique « Bootstrap » produiront les estimations appropriées.

Enfin, il y a un autre problème concernant le changement de ménage. Comme l'objectif de l'ELNEJ est de suivre les enfants, si un enfant longitudinal quitte son ménage initial suite à un divorce dans la famille ou pour d'autres raisons, l'enfant sera suivi dans son nouveau ménage. Le ménage initial est en quelque sorte abandonné. De plus, avec le vieillissement de la cohorte originale, de plus en plus de jeunes quittent leur ménage initial, amplifiant ce phénomène au fil des cycles. Il y a donc des changements de ménages dans le temps, ce qui rend encore plus complexe la création de poids ménages, ou du moins de poids ménages longitudinaux.

Dans bien des cas, la question analytique peut être reformulée pour tenir compte de cette limitation de l'ELNEJ. Par ailleurs, si la question analytique ne peut être reformulée dans une perspective qui part de l'enfant, soit parce que c'est impossible, soit parce que l'objectif même de la recherche est d'étudier les ménages ou une autre unité d'analyse, l'utilisation d'autres sources de données plus appropriées à ce «type» d'analyse devrait être envisagée. Sinon, une des options possibles, serait de ne pas utiliser de pondération, par contre, aucunes conclusions ne pourront être généralisées à l'ensemble de la population. Ici, malgré que nous ne recommandons pas l'utilisation du poids moyen, cette option est « mieux vue » que de ne pas utiliser de poids.

Note d'information

Les Fichiers ICRPS-ELNEJ

Par Cara B. Fedick

Le D[r] J. Douglas Willms et les membres de son personnel à l'Institut canadien de recherche en politiques sociales (ICRPS) de l'Université du Nouveau-Brunswick (campus de Fredericton) ont élaboré un ensemble de fichiers à l'intention des chercheurs qui s'intéressent aux ensembles de données de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) de Statistique Canada. « *Les Fichiers* », qui se composent des données et de la syntaxe de la SSTE, visent à aider les chercheurs à mener des analyses longitudinales à efficacité accrue, au moyen des données de l'ELNEJ.

Les Fichiers constituent une reconfiguration des quatre premiers cycles (le Cycle 1 de 1994-1995; le Cycle 2 de 1996-1997; le Cycle 3 de 1998-1999; et le Cycle 4 de 2000-2001) de l'ELNEJ. *Les Fichiers* comprennent des variables, échelles et mesures dérivées par le D[r] Willms aux fins des analyses employées dans son ouvrage publié *Vulnerable Children : Findings from Canada's National Longitudinal Survey of Children and Youth*[1]. Ces variables supplémentaires, connues comme étant les Variables ICRPS, sont fournies par les chercheurs dans le but de minimiser la redondance de recodage des variables parmi les utilisateurs qui exécutent des analyses semblables. *Les Fichiers* donnent également aux utilisateurs l'occasion de profiter des connaissances et du savoir-faire propres aux techniques d'analyse des chercheurs chevronnés, permettant d'épargner un temps considérable à la rédaction de la syntaxe. De cette façon, les chercheurs peuvent poursuivre leurs analyses et tirer profit de la richesse des données de l'ELNEJ à leur plein potentiel. Même les utilisateurs qui ne s'intéressent pas aux Variables ICRPS pourront profiter de l'utilisation des *Fichiers* afin d'unifier efficacement les fichiers des composantes d'origine (c.-à-d. primaires, secondaires, de garde, etc.) et de mener des procédures exactes de manipulation et d'épuration de données. *Les Fichiers* permettent de reconfigurer l'ELNEJ en un format mieux adapté à l'analyse longitudinale, particulièrement avec le programme *Modélisation hiérarchique linéaire (MHL)*.

Les composantes de la syntaxe dans *Les Fichiers* sont fournies aux utilisateurs à titre de moyen de documentation des Variables ICRPS. Les utilisateurs ayant une bonne connaissance de la syntaxe de la SSTE, peuvent utiliser ces fichiers pour comprendre comment les Variables ICRPS ont été créées ou, dans le cas des utilisateurs de niveau avancé, pour créer des variables semblables à l'aide des autres ensembles de données.

Les Fichiers se composent d'un ensemble de fichiers de données (.sav) et de syntaxe (.sps) de la SSTE, de même que d'un certain nombre de fichiers de directives et de documentation. Le dossier *V1.0 CRISP-NLSCY Files* comporte deux sous-dossiers, 'Data' et

Willms, J. D. (Ed.). (2002). *Vulnerable Children: Findings from Canada's National Longitudinal Survey of Children and Youth*. Edmonton: University of Alberta Press and Human Resources Development Canada, Applied Research Branch.

'Syntaxes and Documentation', de même qu'un document en format .pdf, '*The Files* User's Guide.pdf'.

- *Data*
  Ce dossier comporte quatre fichiers de données de la SPTE (.sav) : ALL_CYCLE1.sav; ALL_CYCLE2.sav; ALL_CYCLE3.sav et ALL_CYCLE4.sav. Ces fichiers contiennent chacun les données des composantes de l'ELNEJ propres à chaque cycle et fournies par Statistique Canada, auxquelles s'ajoute un ensemble de Variables ICRPS supplémentaires. Par exemple, le fichier ALL_CYCLE1.sav comprend les fichiers primaires, secondaires, autoadministrés et de garde du Cycle 1 de l'ELNEJ, de même qu'un recueil des variables du Cycle 1 de l'ELNEJ que les membres du personnel de l'ICRPS ont créé.

- *Syntaxes and Documentation*
  Ce dossier comporte plusieurs sous-dossiers, sauf un qui comporte les syntaxes (.sps files) servant à créer les Variables ICRPS trouvées dans les ensembles de données décrits ci-dessus. Le 'Original StatCan Documentation' constitue l'exception : il contient toute la documentation disponible sur l'ELNEJ (p. ex., les guides de codage, les guides de l'utilisateur, les outils d'enquête, etc.) publiée par Statistique Canada. Les dossiers qui contiennent les syntaxes sont identifiés selon la variable ou l'ensemble de variables, et chaque dossier contient quatre syntaxes, soit une syntaxe pour chacun des cycles d'enquête de la l'ELNEJ.

- *The Files User's Guide.pdf*
  Ce document constitue la principale source de référence des utilisateurs ayant recours aux *Fichiers.* Il comporte les renseignements sur la structure et la préparation des *Fichiers*, de même que les directives sur le comment et le pourquoi de leur utilisation efficace.

Il convient de traiter les composantes des données de la SSTE dans *Les Fichiers* comme tout autre ensemble de données hébergé dans un CDR, puisque le contenu des données des *Fichiers* est dérivé des fichiers protégés de l'ELNEJ. Comme toute autre production de CDR, les analyses menées à l'aide de ces données doivent faire l'objet de demandes de divulgation avant leur autorisation de sortie du CDR. Cependant, les composantes de la syntaxe de la SSTE et le document *The Files User's Guide.pdf* peuvent sortir du CDR sans autorisation.

Les versions à jour des *Fichiers* seront distribuées à tous les CDR (en format cédérom) à intervalles réguliers, au fur et à mesure de la disponibilité des nouvelles parutions de données et des modifications apportées aux *Fichiers*. Les utilisateurs seront informés des mises à jour directement à partir de l'ICRPS.

Les utilisateurs approuvés les plus récents des *Fichiers* sont membres du Réseau des nouveaux chercheurs (RNC). Celui-ci, qui fait partie de l'Institut canadien de recherches avancées (ICRA), a été fondé en 2003 à titre de groupe de longue durée de jeunes chercheurs d'avant-garde dans le domaine du développement humain et cherche à promouvoir la recherche

fondée sur l'ELNEJ. Afin d'avoir accès aux *Fichiers*, ces utilisateurs ont d'abord fait une demande au Conseil de recherches en sciences humaines (CRSH) afin d'avoir accès à l'ELNEJ dans les CDR. Après approbation, ces chercheurs ont dû communiquer avec l'ICRPS pour convenir d'une série de règles et règlements encadrant leur utilisation des *Fichiers*. Ces utilisateurs, ainsi que tous les autres qui obtiendront une autorisation à l'avenir, sont autorisés à accéder aux *Fichiers* quel que soit le CDR au pays.

À la suite de la publication des *Fichiers*, il convient d'informer les utilisateurs potentiels qui souhaitent obtenir l'accès aux *Fichiers* de faire d'abord une demande au CRSH pour obtenir l'accès au CDR puis, une fois cet accès obtenu, communiquer avec l'ICRPS par courriel (CRISPFILES@email.unb.ca) en précisant leur nom, leur affiliation ainsi qu'une description générale de leur besoin d'accéder aux *Fichiers*. L'ICRPS, qui demeure investi du pouvoir d'accorder officiellement la permission et de la responsabilité de traiter les procédures nécessaires donnant aux utilisateurs potentiels l'accès aux *Fichiers*, informera les analystes des CDR des modifications ou ajouts apportés à la liste des utilisateurs autorisés. Chaque analyste de CDR aura la responsabilité de faire en sorte que l'accès aux *Fichiers* au sein de leur CDR soit limité aux utilisateurs autorisés.

Pour obtenir plus de renseignements sur les *Fichiers* ICRPS-ELNEJ, n'hésitez pas à communiquera avec l'Institut canadien de recherche en politiques sociales (ICRPS) de l'Université du Nouveau-Brunswick par courriel au CRISPFILES@email.unb.ca, et à visiter le site (en anglais) http://www.unbcrisp.ca/learningbar/ pour obtenir plus de renseignements sur le projet pour lequel les *Fichiers* ICRPS-ELNEJ ont été conçus à l'origine.

Directives pour les auteurs

Les articles portant sur les questions méthodologiques et les sujets techniques reliés aux données qui se trouvent dans les CDR sont appropriés pour le Bulletin technique et d'information.

Langage du matériel soumis

Les manuscrits peuvent être soumis en français ou en anglais. Une fois accepté, les manuscrits seront traduits dans la deuxième langue officielle avant de les publier.

Longueur d'une soumission

Les articles ne doivent pas dépasser 20 pages à double interligne. Le Bulletin accepte également les notes et les commentaires brefs (idéalement, trois pages ou moins) traitant sur des solutions rapide aux problèmes analytiques soulevées antérieurement dans le Bulletin ou part les chercheurs collègues.

Le format électronique et la mise en page des manuscrits

Les manuscrits doivent être en format "Microsoft Word (.doc)". Les auteurs peuvent les soumettre par courrier ordinaire sur disquette ou disque compact. Ils peuvent également les envoyer comme attachement à un courriel.

Les noms des auteurs, le nom de l'établissement principal, et les coordonnées (numéro de téléphone, adresse postale et adresse électronique) du chercheur principal doivent paraître à la page couverture du manuscrit.

Les auteurs doivent se servir de la police Times New Roman de 12 points, interligne double, et des marges de 1 pouce (2,5 cm) en rédigeant leurs manuscrits.

Nous mettons la majuscule qu'au premier mot du titre (p.e. Pour une utilisation plus conviviale de la méthode bootstrap…).

Nous nous servons des caractères gras que pour les entêtes. Il ne faut pas souligner les mots ou les phrases ni faut il se servir des caractères en italiques pour les entêtes.

Les notes bas de page et les références doivent être à simple interligne. Les auteurs sont invités de consulter *Le guide du rédacteur*, 2^{e} édition.

Le format et mise en page des graphiques et tableaux

Les tableaux et graphiques doivent être soumis en format « Microsoft Excel (.xls) » ou en format séparation par virgule (.csv). Le nom des dossiers doit indiquer le contenu (p.e. tableau1, graphique6, etc.).

Les auteurs peuvent les soumettre par courrier ordinaire sur disquette ou disque compact. Ils peuvent également les envoyer comme attachement à un courriel.

Indiquez dans le texte l'emplacement des tableaux et graphiques plutot que de les placer pas dans le texte. Servez vous du titre suivit par le nom du fichier entre parenthèses. p.e.

Graphique 6. La consommation du chocolat par les enfants au Canada, 2000 (graphique6)

Les expressions mathématiques

Toutes les expressions mathématiques doivent être dissociées du texte. Les équations doivent être numérotées, le numéro devant figurer à la droite de l'équation, aligné à la marge.

Guide de rédaction à l'intention des auteurs

Les auteurs sont priés de se servir de *Le guide du rédacteur*,2e édition. Vous pouvez en acheter une copie du Publications du gouvernement du Canada, Travaux publics et services gouvernementaux Canada.

Où soumettre les manuscrits

Envoyez les manuscrits et toutes communications reliées au Bulletin au Comité de révision.

- Adresse électronique – rdc-cdr@statcan.ca

Révision des soumissions

Le processus de révision initiale des articles relève du Comité de rédaction. Les rédacteurs peuvent inviter des auteurs ayant déjà publié des articles dans le BTI ou des spécialistes à participer au processus. Les articles soumis au Bulletin font l'objet d'une révision permettant d'en assurer l'exactitude, la cohérence et la qualité.

Au terme du processus de révision initiale par le Comité de rédaction, les articles sont soumis à un examen par les pairs et à un examen interne. L'examen par les pairs sera effectué conformément à la Politique concernant l'évaluation des produits d'information de Statistique Canada. En outre, des cadres supérieurs de Statistique Canada procéderont à des examens internes pour s'assurer que le matériel respecte les directives et les normes du Bureau et qu'il ne

porte pas atteinte à la réputation d'impartialité politique, d'objectivité et de neutralité de Statistique Canada.

Veuillez communiquer avec le comité de révision à l'adresse ci haut pour des plus amples renseignements.