Enquête Sur La Santé Dans Les Collectivités Canadiennes (ESCC)
Cycle 2.2 – Nutrition

Documentation sur Software for Intake Distribution Estimation (SIDE)

Statistique Canada

Révisé en Mai 2007

Statistics Canada Statistique Canada

Canada

Table des matières

Chapitre 1 Préface à l'estimation à l'aide du logiciel Software for Intake Distribution Estimation (SIDE)

1.1 Introduction

Un des buts de l'Enquête sur la Santé dans les Collectivités Canadiennes – Nutrition (cycle 2.2) est d'estimer les distributions des apports habituels de différents nutriments à l'échelle provinciale pour quinze groupes d'âge et de sexe. Pour ce faire, on a récolté l'information en effectuant un ou deux rappels de 24 heures de ce que les répondants avaient consommés, c'est-à-dire en effectuant des mesures de leurs apports quotidiens. L'apport quotidien d'un individu est la quantité d'un nutriment qu'il a mangé dans une journée tandis que l'apport habituel peut-être vu comme l'apport quotidien moyen sur une grande période de temps. Pour passer de la distribution des apports quotidiens d'une population à la distribution des apports habituels, il est nécessaire d'ajuster un modèle d'erreur de mesure. Le logiciel SIDE (Software for Intake Distribution Estimation) peut faire ce travail. En général, il est plus intéressant d'étudier la distribution des apports habituels d'une population que la distribution des apports quotidiens. En effet, connaître l'ampleur de la fraction de la population qui a de mauvaises tendances à long terme est plus important que de savoir l'ampleur de la fraction qui a de mauvaises tendances un jour donné.

Pour des estimations qui n'ont pas de lien avec la distribution des apports habituels, il n'est pas nécessaire d'utiliser SIDE. Par exemple, si on veut étudier la distribution des apports quotidiens on peut utiliser simplement le logiciel statistique de son choix.

Il y a trois types d'estimations que l'on peut faire avec la distribution des apports nutritionnels habituels estimée (de la moins complexe à la plus complexe):

1- La moyenne des apports habituels;
2- Le pourcentage de la population qui a un apport habituel sous un certain seuil (« cut-off » en anglais);
3- Un percentile de la distribution.

Sous le modèle d'erreur de mesure ajusté, la moyenne des apports habituels est égale à la moyenne des apports quotidiens. Il n'est donc pas nécessaire d'utiliser SIDE (d'ajuster le modèle) pour obtenir ce type d'estimation. Estimer un pourcentage ou un percentile requiert toutefois l'utilisation du logiciel.

Certaines versions de SIDE permettent l'analyse de la distribution de l'apport habituel en aliments. Ce type d'analyse est différente de l'analyse des distributions de nutriments. Il faut noter que l'analyse d'aliments est possible avec les données de l'Enquête sur la Santé dans les Collectivités Canadiennes - cycle 2.2, mais seulement pour les aliments qui ont une fréquence très élevée dans l'alimentation des individus. Il y avait au maximum deux rappels de 24 heures (deux mesures de l'apport quotidien) par individu lors de la collecte et il en aurait fallu plus pour avoir des résultats suffisamment précis dans les analyses d'aliments moins fréquents.

SIDE utilise la méthode décrite dans Nusser et al. (1996)[1]. Cette méthode repose sur un modèle d'erreur de mesure qui permet de passer de mesures d'apports quotidiens sur les individus à une estimation de la distribution des apports habituels dans la population. La méthodologie du logiciel est divisée en quatre grandes étapes :

1. des ajustements préliminaires;
2. une transformation semiparamétrique pour obtenir la normalité des données;
3. l'estimation des variances inter-individuelle et intra-individuelle des apports quotidiens (ce qui permet l'estimation de la distribution des apports habituels dans l'échelle normale);
4. une transformation de la distribution des apports habituels de l'échelle normale à l'échelle originale.

Statistique Canada recommande l'utilisation de SIDE pour l'estimation des caractéristiques des distributions d'apports habituels à partir des données de l'Enquête sur la Santé dans les Collectivités Canadiennes (ESCC) - cycle 2.2 (volet nutrition). En effet, ce logiciel est à la fois complexe, complet et précis pour ce type d'estimation.

L'utilisation de SIDE est ardue, spécialement quand vient le temps de faire des estimations par intervalles de confiance des caractéristiques d'intérêt. Le guide d'utilisateur de SIDE et la documentation fournie par Statistique Canada devraient permettre aux utilisateurs d'acquérir un bon niveau de confiance pour l'estimation avec SIDE. Le guide d'utilisateur officiel du logiciel (ainsi que le guide technique) est disponible auprès du Department of Statistics and Center for Agricultural and Rural Development de la Iowa State University (ISU). Le guide pour la version SAS/IML de SIDE se trouve à l'adresse :

http://www.card.iastate.edu/publications/DBS/PDFFiles/96tr30.pdf .

Pour plus de renseignement sur les techniques statistiques au cœur du logiciel, on peut consulter le guide technique à l'adresse :

http://www.card.iastate.edu/publications/DBS/PDFFiles/96tr32.pdf .

Finalement, le formulaire pour faire la commande de SIDE peut être téléchargé à l'adresse :

http:/cssm.iastate.edu/software/SIDEodfm.pdf .

Il est important d'utiliser le logiciel de façon correcte et efficace. D'autre part, les particularités des données de l'Enquête sur la Santé dans les Collectivités Canadiennes - cycle 2.2 doivent être prise en compte lors de l'utilisation. Pour arriver à ces fins, on peut suivre les étapes suivantes :

- Bien comprendre les détails de l'enquête en lisant la documentation fournie;

1. Nusser S.M., A.L. Carriquiry, K.W. Dodd and W.A. Fuller (1996). A semiparametric transformation approach to estimating usual intake distributions. *Journal of American Statistical Association*, 91: 1440-1449.

- Être à l'aise avec l'utilisation de la méthode de réplication du bootstrap;
- Terminer de lire ce chapitre afin de choisir la version de SIDE à utiliser. On conseille d'utiliser la version SAS/IML du logiciel. La plupart des recommandations contenues dans la documentation de Statistique Canada concerne cette version;
- Devenir familier avec l'estimation ponctuelle avec SIDE. À cette fin, il faudrait lire le guide d'utilisateur et son complément (le chapitre 2 de ce document) et faire quelques estimations ponctuelles avec SIDE avec différentes valeurs des paramètres, différents domaines d'intérêt et variables étudiées. L'utilisateur voudra bien connaître les entrées et les sorties du logiciel ainsi que les messages d'erreur. De plus, il serait bon de produire quelques graphiques des distributions pour bien comprendre la méthodologie du logiciel;
- Se lancer dans l'estimation par intervalle de confiance en appliquant la méthode de réplication du bootstrap tout en suivant les recommandations contenues dans le dernier chapitre et l'Annexe de ce document.

1.2 Différences entre les versions du logiciel SIDE

AVERTISSEMENT :

Il est possible d'acheter le logiciel SIDE auprès de la Iowa State University (ISU), mais aucun support n'est offert aux usagers. La politique de Statistique Canada est la même**.**

Depuis 1996, l'Iowa State University (ISU) a développé une série de logiciels pour estimer la distribution des apports nutritionnels habituels. Ces logiciels existent en trois versions : SIDE-IML, C-SIDE et PC-SIDE.

SIDE-IML est la version de SIDE qui a été utilisée pour les analyses de Statistique Canada; C-SIDE est la version disponible sous l'environnement UNIX; PC-SIDE est une version qui fonctionne sous l'environnement WINDOWS et qui est encore en développement. Cette section du document relate les principales différences entre ces trois versions du logiciel. L'utilisation de la version SIDE-IML est recommandée pour tous les besoins d'analyse reliés à l'enquête en raison de ses avantages par rapport aux autres versions. Ces avantages seront discutés dans la suite.

1.2.1 SIDE-IML

SIDE-IML est un programme SAS écrit dans le langage matriciel IML. Il consiste en plusieurs modules qui définissent un objet IML qui doit être appelé dans une procédure IML. Pour l'utiliser, on doit d'abord créer des fichiers de données SAS en entrée qui définissent les données et les paramètres à utiliser et ensuite appeler la procédure IML.

Avantages:

1. C'est la version de SIDE la plus malléable. Il est possible d'éditer les modules IML pour adapter le logiciel à des besoins spécifiques, en particulier pour le traitement des répliques bootstrap et l'estimation de la variance échantillonnale.
2. L'utilisateur bénéficie des avantages associés à l'utilisation du logiciel SAS, par exemple pour la création de macros SAS, de graphiques ainsi que pour la manipulation de jeux de données.

Désavantages:

1. Requiert l'utilisation du logiciel SAS, c'est-à-dire que l'utilisateur doit avoir une licence valide du logiciel.
2. L'utilisateur doit être familier avec SAS.
3. Permet l'analyse des nutriments, mais pas des aliments. Ce désavantage est mineur dans cette enquête puisqu'on ne dispose que de un ou deux rappels par répondant, ce qui est trop peu dans la majorité des analyses d'aliments pour avoir des estimations raisonnablement précises.

1.2.2 C-SIDE

C-SIDE est un programme écrit dans le langage C qui doit être exécuté dans un environnement UNIX.

Avantages:

1. Permet l'analyse des distributions des apports en aliments.

Désavantages:

1. L'utilisateur n'a pas autant de liberté du point de vue des fichiers et des graphiques en sortie comparativement à SIDE - IML.
2. Les manipulations graphiques et de jeux de données sont limitées. Plusieurs manipulations de jeux de données doivent être effectuées à l'extérieur du logiciel.

Note :

Les résultats obtenus avec C-SIDE et avec SIDE-IML sont en général légèrement différents (bien que tous deux valides) à cause de petites différences méthodologiques.

1.2.3 PC-SIDE

PC-SIDE est conçu pour fonctionner dans un environnement WINDOWS et consiste en un exécutable relativement convivial. Il n'est recommandé d'utiliser ce logiciel que pour l'exploration de données.

Avantages:

1. PC-SIDE est facile à utiliser pour un usager peu familier avec SIDE.

Désavantages:

1. La réplication bootstrap est très ardue.
2. Il y a très peu de liberté du côté de la manipulation des jeux de données et de la production des graphiques. Les manipulations des données doivent être effectuées à l'extérieur du logiciel.

Chapitre 2 Complément à la version SIDE-IML du guide d'utilisateur de Software for Intake Distribution Estimation (SIDE) pour l'estimation ponctuelle

2.1 Introduction

Dans ce chapitre, l'utilisateur se familiarisera avec l'utilisation de SIDE-IML avec les données de l'ESCC. Il ou elle deviendra plus familier avec la version SIDE-IML disponible en SAS en plus d'être initié aux jeux de données disponibles et aux paramètres et aux options du logiciel. Les messages d'erreur seront aussi discutés.

2.2 Jeux de données à analyser disponibles

Format du fichier contenant l'ensemble des variables nécessaires à l'analyse : Le fichier SAS contenant le jeu de données à analyser doit posséder une variable numérique autant pour identifier les répondants que pour identifier les rappels. Aucune valeur manquante ne peut exister pour les variables d'identification, pour les variables nuisibles (variables des fichiers CLASSVAR et CONTSVAR) ou pour les variables de catégories (variables du fichier BYVAR). Tous les noms de variables doivent être en caractères majuscules et avoir au plus 8 caractères. Puisque certains individus ont 2 rappels alimentaires de 24 heures, le jeu de données en entrée doit contenir 2 enregistrements par individu. Il doit aussi être ordonné par individu et par rappel.

Afin de faciliter la gestion des données pour l'utilisateur, Statistique Canada a créé le fichier HS_SIDE.txt qui lie les variables du fichier maître (HS.txt) aux variables du fichier contenant l'information nutritionnelle des rappels de 24 heures (R24.txt). Ce jeu de données respecte les conditions énoncées au paragraphe précédent. L'information nutritionnelle disponible prend la forme de totaux des apports nutritionnels quotidiens en nutriments pour chaque rappel et chaque individu. Pour l'analyse de sous-totaux (par repas par exemple), l'utilisateur doit créer son propre jeu de données.

Le jeu de données HS_SIDE.txt est décrit plus en détails dans HS_SIDE_lbf.sas et DvDoc_F.doc.

Comme il l'est décrit dans la section IV du guide de l'utilisateur, il faut soumettre les données et les options par l'entremise de plusieurs jeux de données pour utiliser le logiciel SIDE-IML. Les jeux principaux sont : ANALYVAR (variables à analyser), BYVAR (catégories à analyser), CLASSVAR et CONTSVAR (variables nuisibles catégoriques et continues), IDVAR (variables d'identification), WTVAR (variables de poids) et DESC (options des fichiers d'entrée, de sortie et méthodes employées). Certaines options du fichier DESC sont énumérées dans les prochaines sections. Des recommandations sur ces dernières y sont également discutées.

2.3 Exemple de définition des jeux de données

Le code SAS qui suit permet de produire l'estimation de la distribution des apports nutritionnels habituels en énergie des hommes de 31 à 50 ans de l'Île-du-Prince-Édouard. L'identificateur du répondant est conservé dans le jeu de données IDVAR. On supprime l'effet du jour de l'interview en incluant la variable ADMDDD au fichier CLASSVAR. La variable FSDDDEKC dans ANALYVAR spécifie d'étudier l'énergie. On indique au programme d'utiliser le poids principal pour l'estimation en incluant la variable FWGT au jeu de données WTVAR. Il faut noter que la définition des chemins situant la librairie side, le jeu de données principal et son cliché d'article supposent que tous les fichiers sont dans le répertoire C: .

```
/* Définition de la librairie contenant l'objet IML */
LIBNAME side 'c:';

/* Définition du jeu de données à analyser */
DATA ANALYSE;
%let datafid="C:\hs_side.txt";
%include "C:\hs_side_i.sas";
/* On limite l'analyse aux individus du domaine d'intérêt */
IF DHHDDRI=10 AND GEODDHR4=1101;
RUN;

/* Définition des paramètres et des jeux de données en entrée et en
sorties */
DATA DESC;
      INLIB = 'WORK';
      OUTLIB = 'WORK';
      DATASET = 'ANALYSE';
      NDAY = 2;
      EST_DAY1= 'Y';
      EST_ISU = 'Y';
      SAVE_PCT = 'Y';
      SAVE_NPP = 'Y';
      SAVE_SMO = 'N';
;
/* Définition de la variable identifiant le répondant */
DATA IDVAR;
      INPUT NAME $8.;
      CARDS;
SIDEID
;
/* Définition d'une variable nuisible (le jour de l'entrevue) */
DATA CLASSVAR;
      INPUT NAME $8.;
      CARDS;
ADMDDD
;
/* Définition de la variable à analyser */
DATA ANALYVAR;
      INPUT NAME $8.;
      CARDS;
FSDDDEKC
;
```

```
/* Définition du poids de sondage */
DATA WTVAR;
      INPUT NAME $8.;
      CARDS;
FWGT
;
/* Appel de la procédure */
PROC IML;
RESET STORAGE = SIDE.OBJ FW = 7;
LOAD MODULE = (SIDE);
RUN SIDE ('WORK',0);
QUIT;
RUN;
```

2.4 Recommandations pour certaines options

2.4.1 CLASSVAR et CONTSVAR

Dans les fichiers CLASSVAR et CONTSVAR, il faut être prudent dans la liste des variables fournies. Lorsqu'on utilise SIDE-IML, on veut estimer la variance des apports habituels entre les individus en enlevant à la variance totale la portion due à la variabilité des apports **propre** aux individus (consulter le guide d'utilisateur de la version SIDE-IML pour plus de détails). On devrait donc éviter d'inclure une variable dans CLASSVAR ou CONTSVAR qui réduirait l'estimation de la variance **entre** les individus. Par exemple, si on étudie une population d'hommes et de femmes dans une même analyse et qu'on inclut la variable donnant le sexe du répondant dans CLASSVAR, le logiciel éliminera une partie de la variance entre les individus alors que la variance propre aux individus ne sera pas touchée. En effet, il centrera les distributions des apports nutritionnels quotidiens des hommes et des femmes à une valeur commune. La distribution résultante provoquera une sous-estimation de la variance entre les individus et ne permettra pas d'obtenir une distribution des apports habituels convenable.

Il est donc recommandé d'inclure seulement des variables qui auraient pour effet de réduire la variance intra-individuelle (propre aux individus) sans toucher à la variance entre les individus. Une telle variable prend des valeurs différentes d'un rappel à l'autre pour un individu donné. Par exemple, Statistique Canada utilise le jour de la semaine (ADMDDD) ou l'indicateur de fin de semaine (ADMDDFW) dans le jeu de données CLASSVAR.

2.4.2 BYVAR

Il est conseillé de ne pas utiliser le fichier BYVAR lorsqu'on veut étudier plusieurs populations (plus d'un domaine à la fois), mais de plutôt faire des analyses séparées pour chacune des populations d'intérêt. Le logiciel cessera de fonctionner s'il rencontre un problème sérieux avec un groupe donné, ce qui aura pour effet que les groupes suivants de la liste ne seront pas traités. Ces derniers ne sont pas nécessairement problématiques. On peut deviner que ce problème prend de l'ampleur quand on veut estimer la variance échantillonnale à l'aide des poids bootstrap.

Lorsque le fichier BYVAR est tout de même utilisé, il est conseillé aux utilisateurs d'utiliser des variables numériques numérotées de 1 au nombre de populations étudiées. SIDE transformera automatiquement les variables alpha-numériques en variables numériques avec une telle numérotation. De plus, un message d'avertissement apparaîtra dans le journal de SAS (le « log » en anglais) donnant l'équivalence entre les deux types de variables. Si la variable indiquant le domaine étudié est numérique mais que ses valeurs ne vont pas de 1 au nombre de domaines étudié, certains fichiers de sortie auront une variable indiquant le groupe d'appartenance (ou le « BY-group », _INT_BY_) numérotée à partir de 1 alors que d'autres auront la variable d'origine. Cette particularité peut entraîner une certaine confusion. Le meilleur moyen d'éviter cette confusion est bien évidemment de ne pas utiliser le fichier BYVAR.

2.4.3 DESC

Ce jeu de données contient les options concernant les fichiers d'entrée, de sortie et les méthodes employées. Une description de ces options est disponible dans le guide de l'utilisateur. Statistique Canada recommande d'utiliser les options suivantes :

- **INLIB, OUTLIB, DATASET** pour définir les répertoires contenant les fichiers d'entrée et de sortie ainsi que le jeu de données en entrée;
- **NDAY = 2** (pour spécifier le nombre de rappels maximum par individu);
- **EST_DAY1= 'Y'**;
- **EST_ISU = 'Y'**;
- **SAVE_PCT = 'Y'**;
- **SAVE_NPP = 'Y'**;

Les quatre dernières options permettent d'obtenir les fichiers PCT1, PCTU, NPP1, NPPU en sortie. Ces fichiers sont particulièrement utiles pour créer les outils de diagnostics graphiques entre autres.

- **NPTS = *n***; (où *n* est un entier plus grand que zéro)

La principale utilité de cette option est qu'on contrôle le nombre d'estimations de quantiles qui sont produites. Si on fixe NPTS à 9999, on obtient toutes les estimations de quantiles de 0.0001 à 0.9999 (0.0001, 0.0002, ..., 0.9998, 0.9999). Une autre utilité est que l'option permet de changer la précision des graphiques produits. Accroître NPTS augmente la précision des graphiques. Par défaut NPTS est égal à 41.

Il est déconseillé d'utiliser directement les apports habituels individuels prédits. Les options suivantes spécifient au logiciel de ne pas créer le jeu de données SMOOTH, un jeu de données qui peut être volumineux et qui contient les prédictions des apports habituels pour chaque répondant.

- **SAVE_SMO = 'N'**;
- **INDIVUI = 'N'**;

Le jeu de données DESC contient également les options **LINFRAC, MAXJP, ADALPHA** qui permettent de solutionner certains types d'erreur.

2.5 Options non documentées dans le guide de l'utilisateur

Certaines modifications ont été apportées au logiciel SIDE-IML depuis la publication du guide de l'utilisateur et certaines options utiles ne sont pas documentées. Par défaut, certains renseignements sont disponibles seulement dans le journal et la fenêtre de sortie de SAS (les fenêtres « log » et « output »). Il peut être intéressant de modifier les modules IML pour sauvegarder ces renseignements dans des fichiers de sortie.

2.5.1 PEVCR et NPEVCR

Deux jeux de données qu'il est possible de fournir à la procédure IML sont omis du guide d'utilisateur : PEVCR et NPEVCR. Les deux fichiers permettent de fournir à la procédure des valeurs forcées de la variance de l'erreur de mesure (variance intra-individuelle) et du quatrième moment centré de l'erreur de mesure. Voici en exemple la définition de ces deux jeux de données où la variance d'erreur de mesure est forcée à 0.75 et le quatrième moment centré de l'erreur de mesure est forcé à 3 :

```
DATA PEVCR;
INPUT NAME;
CARDS;
0.75
3.00
;
RUN;

DATA NPEVCR;
INPUT NAME;
CARDS;
9999
;
RUN;
```

Le jeu de données NPEVCR spécifie l'importance relative à accorder aux valeurs forcées par rapport à l'estimation obtenue avec les données. La valeur 9999 indique au logiciel de donner tout le poids à la valeur forcée et donc de ne pas considérer la valeur estimée à l'aide des données. Ces jeux de données sont particulièrement pratiques lorsqu'on n'obtient pas du logiciel les estimés des deux paramètres (on en parle plus en détail dans la prochaine section qui porte sur les messages d'erreur). D'ailleurs, le jeu de données VARCOMP en

sortie contient toute l'information sur les estimations de variances. Les observations 4 et 6 contiennent respectivement la variance d'erreur de mesure et le quatrième moment centré de l'erreur de mesure. Le guide d'utilisateur contient la description des 19 autres observations du fichier VARCOMP.

2.5.2 CDFU1 (fichier de sortie supplémentaire)

Par défaut, le logiciel imprime dans la fenêtre de sortie de SAS les estimés des probabilités relatives à la distribution de l'apport habituel qu'on a spécifié au préalable dans le fichier CUTOFFS. Pour qu'un fichier de ces probabilités spécifiques soit produit en plus d'avoir les résultats dans la fenêtre de sortie, on peut ajouter les lignes suivantes à la fin du module STORDATA du code source IML :

```
START STORDATA(_G_,OUTLIB,NAMES);
RESET STORAGE = WORK.STRG;

[.....................]

            LOAD CDFU1;
          CALL EXECUTE('CREATE WORK.CDFU1 FROM CDFU1[COLNAME = {"Value"
"Prob Below" "Prob Above" "Std Error"}];');
              APPEND FROM CDFU1;
              CALL EXECUTE('CLOSE WORK.CDFU1;');
              FREE CDFU1;
FINISH;
store module = (stordata);
free module = (stordata);
```

Il faut noter que ces modifications produiront un message d'erreur lorsque SIDE est exécuté sans qu'un jeu de données CUTOFF ne soit défini a priori. Le programme modifié tente de créer le jeu de données WORK.CDFU1 à partir de l'objet IML CDFU1 qui n'existe pas lorsque aucun jeu de données CUTOFF n'est défini. Le message d'erreur n'a aucun impact sur les résultats et peut être ignoré.

2.5.3 WARNING (fichier de sortie supplémentaire)

Afin d'avoir les messages d'erreur dans un fichier SAS en plus de les avoir dans le journal (ou le « log »), le module WARN du code IML de SIDE pourrait être changé pour :

```
START WARN(ERRCODE);
TMP = CHAR(ERRCODE,3);
CALL CHANGE(TMP,' ','0',0);
MSG = CONCAT('*** SIDE WARNING #',TMP,' ***');

/* Partie modifiée */
call execute ('WARN="000";',
'use work.WARNING;',
'read all;',
'close work.WARNING;',
```

```
'if WARN = "000" then WARN=TMP;',
'else WARN=WARN//TMP;',
'tempor=concat(WARN,TMP);',
'create work.warning from WARN[COLNAME = {"WARN"}];',
'append from WARN;',
'close work.warning;');
/* Fin de la partie modifiée */

PRINT MSG;
FREE TMP L MSG;
FINISH;
store module = (warn);
free module = (warn);
```

Il faut aussi modifier le début du module SIDE de la façon suivante pour que le fichier WARNING soit créé en sortie:

```
START SIDE(DESCLIB,QUIET);
RESET NONAME CENTER LOG;
CALL EXECUTE("OPTIONS NONOTES;TITLE;");
PRINT / 'SIDE/IML Version 1.11';
PRINT 'Copyright 2001';
PRINT 'Iowa State University Statistical Laboratory';
PRINT 'All rights reserved';
PRINT ,'Author: Kevin W. Dodd',;

PRINT '**** RUNNING ****',,,;
RESET NOCENTER;
RESET STORAGE = side.obj;
LOAD MODULE = (WARN _VTYPE_ READ_KW PROC_DS CHECK_G_ READDATA
               READ_DS DMP_PID PROC_BY SETUP SQZ INIT_DS);
RESET CENTER;
PRINT 'Running Initialization Procedure...';
RESET NOCENTER;

/* Fichier warning */

create work.warning;
close work.warning;
[.............................]
```

2.6 Messages d'erreur

SIDE indique des erreurs potentielles à l'aide de messages d'avertissement numérotés apparaissant dans le journal (le « log »). La liste et la description détaillée de ces avertissements sont disponibles dans la section VIII (SIDE Warning Codes) du guide de l'utilisateur.

La plupart des messages d'erreur qui apparaissent dans le journal de SAS indique des problèmes avec les données en entrée ou avec les paramètres. Généralement, ces messages sont fatals, c'est-à-dire que SIDE cesse de fonctionner. Certains messages ne sont pas fatals et SIDE complètera alors sa procédure d'estimation. Il est donc important de s'assurer de la pertinence de ces messages d'erreur.

Les messages les plus courants liés à des problèmes dans l'estimation sont les messages 61 (dérivées négatives dans le polynôme de transformation à la normalité), 63 (la transformation fait qu'on rejette l'hypothèse de normalité) et 65 (la variance estimée des apports habituels est négative). Parmi ces trois messages, seul le 65 est fatal. Il est important de corriger toutes ces erreurs parce que les messages indiquent qu'une hypothèse de base du modèle pourrait ne pas être respectée. Les messages d'erreur 61 et 63 peuvent revenir à plus d'une reprise dans une même application de la procédure parce que le logiciel tente plusieurs transformations. Il y a un véritable problème de type 61 et/ou 63 lorsque toutes les transformations ont un message d'erreur.

Parmi les solutions à envisager pour corriger les erreurs il y a :

- Vérifier les données fournies au logiciel et les sorties à chaque étape à l'aide d'outils graphiques pour cibler la raison du problème;
- Forcer le paramètre **LINFRAC** à une valeur plus élevée. Cependant, il n'est pas recommandé d'aller au delà de 5%. Changer ce paramètre semble corriger la plus grande partie des erreurs reliées aux avertissements 61 et 63;
- Étudier une population cible plus grande afin d'améliorer la précision des estimations. La taille d'échantillon sera alors augmentée et la probabilité d'avoir une estimation de variance négative (message numéro 65) sera alors diminuée. Le calcul de cette probabilité est discuté dans le chapitre suivant;
- Dans de rares cas, il sera bon de diviser la population étudiée en plusieurs parties et d'étudier chacune de ces parties séparément. Cela peut être approprié entre autres si on a une distribution avec plus d'un mode, donc très difficile à rendre normale. Séparer la population judicieusement pourrait alors permettre d'avoir plusieurs populations unimodales à étudier séparément;
- Si on a un message numéro 65, on peut forcer la variance de l'erreur de mesure ainsi que le quatrième moment centré de cette dernière à l'aide de valeurs externes à l'aide des jeux de données PEVCR et NPEVCR.

Certaines solutions aux erreurs sont décrites plus en détails dans chapitre 3 intitulé : « Mesure de la Variabilité d'Échantillonnage à l'Aide de Répliques Bootstrap avec Software for Intake Distribution Estimation (SIDE) ». En effet, les erreurs ne sont pas

propres à l'estimation ponctuelle mais surviennent aussi dans l'estimation pour une réplique bootstrap. Le traitement de ces erreurs se trouve alors compliqué. Les principaux problèmes qu'on rencontre avec l'estimation bootstrap sont reliés avec les messages d'erreur 65 (estimation de variance inter-individuelle négative). Il peut arriver que le poids principal donne une estimation, mais qu'un nombre de répliques bootstrap ne donne pas d'estimation. Ainsi, la solution qu'on choisit pour le poids principal ne sera pas nécessairement suffisante pour obtenir toutes les estimations bootstrap. Il est alors plus approprié de choisir une solution qui résoudra les problèmes rencontrés avec le poids principal et avec l'ensemble des estimations bootstrap. Toutes ces méthodes sont décrites en détails dans le chapitre suivant.

Chapitre 3 Mesure de la variabilité d'échantillonnage à l'aide de répliques Bootstrap avec Software for Intake Distribution Estimation (SIDE)

3.1 Introduction

Pour des enquêtes avec des plans de sondage simples (par exemple des plans aléatoires simples ou stratifiés), il existe des formules mathématiques permettant d'estimer la variance échantillonnale. L'Enquête sur la Santé dans les Collectivités Canadiennes (ESCC) a un plan de sondage complexe ce qui implique qu'il n'existe pas de formule mathématique pour calculer la variabilité d'échantillonnage. Il est alors nécessaire d'utiliser une méthode de réplication de l'échantillon pour estimer cette variance et la méthode la plus appropriée est celle du bootstrap. BOOTVAR est un programme disponible dans les langages SAS et SPSS qui a été développé par Statistique Canada pour estimer la variance échantillonnale à partir de répliques bootstrap pour des enquêtes à plan de sondage complexe comme l'ESCC.

Pour des estimations simples comme des totaux, des ratios ou des paramètres de régressions, BOOTVAR permet d'estimer la variabilité échantillonnale à partir du fichier de poids bootstrap. Afin de faire cette estimation, la macro calcule le paramètre d'intérêt (total, ratio, ...) une fois par réplique et calcule la variance entre les 500 valeurs obtenues. Pour des estimations reliées à la distribution des apports habituels obtenues avec SIDE, il faut imiter ce processus. Il faut donc estimer les paramètres d'intérêt avec SIDE une fois par réplique (par poids bootstrap) et calculer la variance entre les 500 estimations obtenues.

Le temps d'exécution de SIDE pour une estimation avec une réplique donnée est plus grand que celui requis pour une estimation simple qu'on ferait avec BOOTVAR. D'ailleurs, le temps total pour les exécutions des 500 répliques avec SIDE peut être très long. En conséquence, il est important de faire les estimations bootstrap de la façon la plus efficace possible.

D'autre part, avoir une estimation principale sans erreur ne signifie pas que toutes les répliques bootstrap n'auront aucune erreur. En effet, la complexité de la procédure de SIDE vient compliquer l'estimation de la variabilité échantillonnale parce que ces erreurs peuvent survenir pour n'importe laquelle des répliques bootstrap. Il faudra donc régler ces problèmes de façon efficace et de la manière la plus automatique possible.

Des conseils généraux sur l'estimation bootstrap suivent en 3.2. La section 3.3 décrit les étapes à suivre pour effectuer l'estimation de la variabilité d'échantillonnage. Vient ensuite en 3.4 une section décrivant une technique pour estimer le nombre de répliques qui auront une erreur d'estimation fatale due à une estimation de variance négative (message d'erreur numéro 65). On donnera bien entendu la liste des techniques pour venir à bout de ces échecs en 3.5 ainsi que les qualités et les défauts de ces méthodes. La dernière section du chapitre donne les temps d'exécution et les choix qui ont été faits pour les analyses de Statistique Canada spécifiques au premier article et à la première série de tableaux

CANSIM. Finalement, l'Annexe donne des descriptions détaillées des méthodes énumérées à la section 3.5.

3.2 Quelques conseils relatifs à l'estimation de la variabilité d'échantillonnage

- Il est recommandé d'avoir des estimations bootstrap pour le plus de répliques possible pour éviter d'avoir une estimation de la variabilité d'échantillonnale biaisée.
- Il est conseillé de créer une boucle dans une macro SAS et d'utiliser un poids à la fois en redéfinissant le fichier WTVAR à chaque itération. Il est possible de faire toutes les estimations bootstrap dans une seule exécution de SIDE en mettant la liste des 500 poids bootstrap dans WTVAR et en répétant la même variable plusieurs fois dans ANALYVAR. Cependant, le temps de calcul sera beaucoup plus long que si on avait utilisé une boucle.
- Si on n'élimine pas les individus d'une réplique donnée qui ont des poids nuls, ils se retrouveront dans le fichier SMOOTH, qui contient les valeurs prédites des apports habituels des répondants, alors qu'ils ne devraient pas y être.
- La plupart du temps, on peut ignorer les messages d'erreur 61 et 63 parce que le logiciel aura réussi à trouver une transformation alternative qui respecte les hypothèses du modèle correspondantes. En effet, le logiciel essaie un bon nombre de transformations qui ne sont pas nécessairement toutes problématiques. On peut donc ignorer les messages 61 et 63 à moins que toutes les transformations aient un message d'erreur.
- On ne peut pas ignorer les messages d'erreur 65 parce que le logiciel ne fournit pas d'estimation ponctuelle pour une réplique qui a un tel message. Il faut donc essayer de trouver une solution (un compromis). Ce document contient une description des différentes solutions possibles avec leurs avantages et leurs désavantages.

3.3 Étapes à suivre pour l'estimation Bootstrap

Voici les étapes suggérées pour faire l'estimation de la variabilité d'échantillonnage à l'aide de répliques bootstrap :

1. Choisir un domaine et une caractéristique d'intérêt;
2. Faire l'estimation avec le poids principal et régler les erreurs s'il y en a en suivant les conseils du chapitre précédent, « Complément à la Version SIDE-IML du Guide d'Utilisateur de Software for Intake Distribution Estimation (SIDE) pour l'Estimation Ponctuelle »;
3. Estimer le nombre de répliques bootstrap qui échoueront avec une erreur de type 65 (variance entre les individus négative) à l'aide des résultats obtenus en 2 et de la technique décrite en 3.4;
4. Choisir la solution aux erreurs qui convient le mieux parmi la liste en 3.5. La même solution sera appliquée à chaque réplique bootstrap et au poids principal;
5. Faire l'estimation bootstrap en utilisant la méthode choisie en 4 en suivant la description dans la section correspondante de l'Annexe.

3.4 Estimation du nombre de répliques qui échoueront

Estimer le nombre de répliques bootstrap qui auront un échec de type 65 avant de calculer les estimations bootstrap peut aider à sauver beaucoup de temps de calcul. En effet, savoir ce nombre à l'avance permet de guider l'utilisateur quant au choix du domaine à analyser et de la méthode à utiliser pour régler ces erreurs.

Si le plan d'échantillonnage de l'enquête était un plan aléatoire simple, on pourrait estimer le nombre de répliques avec une erreur fatale reliée à une estimation de variance entre les individus négative à partir des résultats obtenus avec le poids principal et à partir de la formule suivante :

$$\text{Nombre estimé de répliques qui échoueront} = 500 \times P\left(F_{b-1}^{a-1} \le \frac{1}{1 + \frac{n_0 \hat{\sigma}_x^2}{(a-1)\hat{\mu}_A}} \right)$$

où :

a = nombre de premier rappels
b = nombre de deuxième rappels
$n_0 = a + b - \frac{1}{a+b}(4b + a)$
F_{b-1}^{a-1} = statistique de loi F à $a-1$ degrés de liberté au numérateur et $b-1$ degrés de liberté au dénominateur
$\hat{\sigma}_x^2$ = variance de l'apport habituel (« usual intake ») estimé avec le poids principal (troisième élément du jeu de données VARCOMP)
$\hat{\mu}_A$ = variance de l'erreur de mesure estimé avec le poids principal (quatrième élément du jeu de données VARCOMP)

Le plan de l'Enquête sur la Santé dans les Collectivités Canadiennes - cycle 2.2 n'est pas aléatoire simple, il est très complexe. Cependant, la formule donnée plus haut peut donner une idée du nombre de répliques qui échoueront sous ce plan. On cherchera à avoir le plus grand nombre de premier et de deuxième rappels dans une analyse pour faire diminuer la probabilité qu'une réplique donnée échoue.

Dans la suite, on étudiera l'apport nutritionnel habituel en énergie des hommes de 31 à 50 ans de l'Île-du-Prince-Édouard. Pour cette population, on a dans l'échantillon (avec ADMDDD le jour de l'interview comme variable nuisible) :

$a = 116$
$b = 55$
$n_0 = 169.04$
$\hat{\sigma}_x^2 = 0.48801$
$\hat{\mu}_A = 0.52363$

$$500 \times P\left(F_{b-1}^{a-1} \le \frac{1}{1 + \frac{n_0 \hat{\sigma}_x^2}{(a-1)\hat{\mu}_A}} \right) = 500 \times P\left(F_{54}^{115} \le 0.4220\right) = 500 \times 0.0001 = 0.05$$

On estime donc le nombre de répliques bootstrap qui échoueront à 0.05. Avec les données, une seule réplique échouera au total, ce qui est vraiment très proche de l'estimation.

Estimer le nombre de répliques qui échoueront est utile quand vient le temps de faire le choix d'une solution à adopter pour régler ces échecs. Cette estimation nous donne une idée de l'ampleur du problème.

3.5 Solutions aux répliques qui échouent

Il peut être difficile d'obtenir des estimations bootstrap pour chacune des répliques, surtout à cause des erreurs de type 65. Les dernières études de Statistique Canada ont permis de trouver plusieurs solutions possibles.

Sous l'hypothèse que le modèle d'erreur de mesure sous-jacent à SIDE est bon, avoir un message 65 avec le poids principal ou avec des répliques bootstrap serait le fruit du hasard (de la malchance). Ceci ne signifie pas pour autant qu'il faut prendre ces messages d'erreur à la légère car ça aurait pour effet de biaiser l'estimation de la variance échantillonnale. Il est donc recommandé, quand une telle chose se produit, d'utiliser une des approches de la liste qui suit :

1. Modifier le domaine d'analyse;
2. Modifier les paramètres et les options en entrée;
3. Ne rien faire;
4. Modifier l'estimateur de variance inter-individuelle pour qu'il prenne des valeurs positives ou nulles et appliquer la méthode de Woodruff pour les estimations de percentiles;
5. Forcer les paramètres relatifs à l'erreur de mesure avec des valeurs différentes pour chaque réplique;
6. Forcer les paramètres relatifs à l'erreur de mesure avec les mêmes valeurs pour toutes les répliques.

Pour aider à choisir la méthode qui convient le mieux aux besoins, le Tableau 1 résume les qualités de chaque approche. De plus, l'Annexe donne une description précise de chacune des méthodes. Il est important de mentionner que pour que l'estimation de la variance échantillonnale d'une estimation ponctuelle obtenue avec le poids principal soit valide, il faut répéter la même estimation avec les répliques bootstrap dans les mêmes conditions. Il faut aussi noter qu'obtenir un grand nombre de répliques bootstrap qui échouent peut être un signe que le modèle ajusté par SIDE n'est pas approprié.

Avant d'entrer dans les détails, il est bon de rappeler qu'il y a trois types d'estimations que l'on peut faire avec la distribution des apports nutritionnels habituels estimée (de la moins complexe à la plus complexe):

1. La moyenne des apports habituels;
2. Le pourcentage de la population qui a un apport habituel sous un certain seuil (« cut-off » en anglais);
3. Un percentile de la distribution.

Dans le Tableau 1, la dernière colonne compare la longueur de l'intervalle de confiance obtenu avec la méthode 4 à la longueur de l'intervalle de confiance de la ligne correspondante. La méthode 4 a été utilisée comme référence parce que l'hypothèse qui lui est sous-jacente est la moins forte dans la plupart des analyses.

Tableau 1 Qualités des approches pour améliorer l'estimation de la variabilité d'échantillonnage

Méthode	Rapide	Simple	Aucune hypothèse additionnelle à faire pour que la méthode soit valide	Longueur de l'intervalle de confiance
1. Modifier le domaine analysé		X	X	Plus petite si le domaine est élargi
2. Modifier les paramètres et les options en entrée		\	X	Comparable
3. Ne rien faire	X	X		Plus petite
4. Estimateur de variance positif	X			**Longueur de référence**
5. Donneur-receveur, réplique par réplique				Plus petite
6. Donneur-receveur, valeur unique	X			Beaucoup plus petite

X = Répond complètement au critère
\ = Répond partiellement au critère

Voici un résumé détaillé des qualités et défauts des approches :

1. Modifier le domaine d'analyse

Description sommaire	• Dans certaines situations, le domaine d'intérêt peut avoir trop peu d'observations pour estimer les paramètres d'intérêt. Quand c'est nécessaire, l'utilisateur devrait choisir un domaine d'analyse moins précis.
Temps d'exécution	• Si on élargi le domaine, le temps d'exécution sera plus long que celui nécessaire pour produire les estimations pour le domaine initial.
Complexité	• Faible.
Hypothèse(s)	• Aucune.
Condition(s) technique(s) d'utilisation	• Le nouveau domaine d'étude ne doit pas avoir de problèmes avec les répliques bootstrap qui échouent.
Utilisation suggérée	• Si l'analyste est prêt à changer de domaine cible.

2. Modifier les paramètres et les options en entrée

Description sommaire	• Dans certaines situations, changer les paramètres en entrée et les options peut résoudre le problème.
Temps d'exécution	• Accru, car il faut tester les modifications.
Complexité	• Moyenne. On doit connaître les détails de la procédure de SIDE pour changer adéquatement les options et les paramètres.
Hypothèse(s)	• Aucune.
Condition(s) technique(s) d'utilisation	• Il doit exister une combinaison des paramètres et options qui résout les problèmes de la grande majorité des répliques. Il faut utiliser ces nouvelles valeurs pour l'estimation ponctuelle (le poids principal).
Utilisation suggérée	• Si l'utilisateur est prêt à prendre le temps et si le nombre de répliques qui échouent est plutôt faible alors cette méthode peut être utilisée. En effet, si ce nombre est élevé, il y a peu de chances qu'assouplir le modèle règle les problèmes de toutes les répliques. Il est à noter qu'il n'est pas garanti que les changements règleront ces problèmes, ce qui peut être considéré comme une perte de temps.

3. Ne rien faire

Description sommaire	• Cette méthode consiste à estimer la variance échantillonnale avec les répliques qui n'ont pas de message d'erreur 65 seulement.
Temps d'exécution	• **Temps de référence.**
Complexité	• Faible.
Hypothèse(s)	• Le biais créé sur l'estimation finale de la variance échantillonnale par les répliques bootstrap supprimées est négligeable. • L'estimation de la variance bootstrap est stable avec le nombre de répliques réduit. En d'autres mots, l'estimation de la variance échantillonnale a convergé vers la vraie valeur de façon satisfaisante.
Condition(s) technique(s) d'utilisation	• L'estimation avec le poids principal doit fonctionner pour avoir une valeur de l'estimation ponctuelle (le centre de l'intervalle de confiance).
Utilisation suggérée	• Si le nombre de répliques qui échoue est très faible car dans ce cas les deux hypothèses sont plus facilement justifiables.

4. Modifier l'estimateur de variance inter-individuelle pour qu'il prenne des valeurs positives ou nulles et utiliser la méthode de Woodruff pour les estimations de percentiles

Description sommaire	• Cette méthode consiste à estimer la variance inter-individuelle en prenant le maximum entre 0 et la valeur donnée par SIDE (qui peut être négative) et à utiliser la méthode développée par Woodruff[2] pour l'estimation de percentiles.
Temps d'exécution	• Le temps d'exécution reste le même que si on ne faisait rien pour corriger les répliques bootstrap, mais le temps de programmation peut être long.
Complexité	• Haute. Il faut bien visualiser chaque étape de l'approche. Il faut en plus comprendre la méthode de Woodruff si on estime des percentiles.
Hypothèse(s)	• Lorsqu'on estime des percentiles avec la méthode de Woodruff, l'utilisateur fait l'hypothèse supplémentaire que la distribution échantillonnale d'un pourcentage calculé à partir de la distribution des apports habituels estimée est approximativement normale. C'est à peu près équivalent à dire que les 500 estimations bootstrap de probabilité d'être sous un seuil donné (« cut-off ») suivent une loi normale.
Condition(s) technique(s) d'utilisation	• Aucune si on est intéressé à estimer des probabilités. • Si on est intéressé à estimer des percentiles, il faut que l'estimation avec le poids principal fonctionne.

2. Woodruff R.S. (1952). Confidence intervals for medians and other position measures. *Journal of American Statistical Association*, 57: 622-627.

Utilisation suggérée	• Son utilisation est toujours recommandée pour l'estimation de probabilités car c'est la méthode avec l'hypothèse la moins forte pour ce type d'estimation. • Le nouvel estimateur de variance, qui est le maximum entre 0 et la valeur donnée par SIDE, est biaisé contrairement à celui utilisé par SIDE. Cependant, l'intervalle de confiance obtenu est généralement moins biaisé que si on avait ignoré les répliques échouant. • Pour l'estimation de percentiles, on suggère d'utiliser cette méthode si le nombre d'estimations à produire est grand parce que c'est une méthode rapide (une fois la programmation réalisée) et c'est une méthode facilement automatisable. • Le point faible de cette méthode est que pour l'estimation de percentiles il faut s'assurer que l'hypothèse soit respectée. On peut le faire en comparant des histogrammes des estimations bootstrap de probabilités à une loi normale ou en faisant des tests de normalité sur ces estimations bootstrap. Règle générale, l'hypothèse aura tendance à ne pas être respectée si le nombre de répliques qui échouent est grand, ce qui correspondra le plus souvent au non respect de la deuxième condition technique (une estimation avec le poids principal qui échoue).

5. Forcer la valeur des paramètres relatifs à l'erreur de mesure réplique par réplique

Description sommaire	• Cette méthode consiste à donner à SIDE des valeurs des paramètres qui sont à l'origine des messages d'erreur 65 au lieu de laisser le logiciel les calculer. Pour tenir compte de l'imprécision dans l'estimation de ces paramètres avec le domaine donneur, on force des valeurs différentes à chaque réplique. Le processus est fait réplique par réplique.
Temps d'exécution	• Au moins doublé comparativement à ignorer les répliques qui échouent.
Complexité	• Haute, à cause de la programmation.
Hypothèse(s)	• Les paramètres de l'erreur de mesure dans le domaine donneur sont les mêmes que ceux dans le domaine receveur. Ceci implique que les différences de consommation entre les individus du groupe donneur sont semblables aux différences de consommation entre les individus du groupe receveur.
Condition(s) technique(s) d'utilisation	• Le domaine donneur ne doit pas avoir de problèmes avec les répliques bootstrap qui échouent (ou très peu). • Le domaine donneur doit provenir de l'enquête pour avoir l'estimation avec le poids principal et les 500 estimations bootstrap de la variance et du quatrième moment centré de l'erreur de mesure.
Utilisation suggérée	• S'il existe un domaine donneur qui pourrait satisfaire l'hypothèse.

6. Forcer la valeur des paramètres relatifs à l'erreur de mesure avec les mêmes valeurs pour toutes les répliques

Description sommaire	• Cette méthode est semblable à la méthode 5, mais la même valeur des paramètres est forcée pour chaque réplique. Ceci revient à négliger l'imprécision qui existe dans l'estimation de ces paramètres avec le domaine donneur.
Temps d'exécution	• Reste le même que si on n'avait rien fait.
Complexité	• Moyenne.
Hypothèse(s)	• Les paramètres de l'erreur de mesure du domaine donneur sont les mêmes que ceux du domaine receveur. • Les paramètres à forcer ont été mesurés avec une précision échantillonnale si grande dans le domaine donneur que l'effet de leur erreur échantillonnale sur la longueur de l'intervalle de confiance produit est négligeable.
Condition(s) technique(s) d'utilisation	• On doit avoir des valeurs à imputer, soit parce qu'on les a calculées, soit parce qu'un expert les a fournies.
Utilisation suggérée	• En dernier recours, si on ne peut utiliser aucune des autres méthodes. Les hypothèses sont très difficilement justifiable.

Voici finalement l'utilisation suggérée de chacune des méthodes en fonction du nombre de répliques qui échouent :

Peu importe le nombre de répliques qui échouent	• 1. Modifier le domaine d'analyse.
Nombre de répliques échouant faible (moins de 5% des répliques)	• 2. Rendre le modèle ajusté par SIDE moins strict en modifiant les paramètres et les options en entrée. • 3. Ne rien faire.
Nombre de répliques échouant moyen (moins de 250)	• 4. Modifier l'estimateur de variance inter-individuelle pour qu'il prenne des valeurs positives ou nulles et utiliser la méthode de Woodruff pour les estimations de percentiles.
Nombre de répliques échouant élevé (250 ou plus) ou s'il n'y a pas d'estimation ponctuelle possible	• 5. Forcer les paramètres relatifs à l'erreur de mesure avec des valeurs différentes à chaque réplique. • (en dernier recours) 6. Forcer les paramètres relatifs à l'erreur de mesure avec les mêmes valeurs pour toutes les répliques.

3.6 Détails de l'exécution des analyses produites pour le premier article de Statistique Canada et pour les premiers tableaux CANSIM

Cette section du document donne les temps de calcul nécessaires à la production de la majeure partie des premiers résultats publiés par Statistique Canada et les approches qui ont été choisies pour traiter les messages d'erreur lors de l'estimation de la variance échantillonnale à l'aide de répliques bootstrap. La majorité des analyses a été faite pour une multitude de domaines d'intérêt sur des distributions de l'apport habituel de variables relatives à l'énergie et sur des distributions de l'apport habituel en aliments très fréquents. Pour les analyses d'aliments, on étudiait les distributions des apports habituels du nombre de portions consommées pour chacun des quatre groupes alimentaires.

Les multiples domaines qui ont été étudiés sont définis par les variables suivantes :

1. DHHDDRI : 13 plus vieux groupes d'âge et de sexe cibles de l'enquête (sexes confondus : 1-3 ans, 4-8 ans, 9-13 ans; par sexe : 14-18 ans, 19-30 ans, 31-50 ans, 51-70 ans et 71 ans et plus);
2. AGE1 : deux classes d'âge : 4 à 18 ans et 19 ans et plus;
3. INCOME : 8 groupes formés du croisement de la variable INCDDIA5 (où les niveaux bas et bas-moyen sont combinés) et de la variable AGE1;
4. AGE_REG : 10 groupes formés du croisement des cinq régions (Atlantique, Québec, Ontario, Prairies et Colombie-Britannique) et de la variable AGE1;
5. DIG2 : 11 groupes d'âge et de sexe propres à l'analyse des produits laitiers (4-9 ans sexes confondus; par sexe : 10-16 ans, 17-30 ans, 31-50 ans, 51-70 ans et 71 ans et plus);
6. AGE2 : trois classes d'âge : 4 à 9 ans, 10 à 16 ans et 17 ans et plus;
7. INC_LAIT : 12 groupes formés du croisement de la variable INCDDIA5 (où les niveaux bas et bas-moyen sont combinés) et de la variable AGE2;
8. AGE2_REG : 15 groupes formés du croisement des cinq régions canadiennes et de la variable AGE2;
9. DIG_PRV : 130 groupes formés du croisement de la province et de DHHDDDRI.

Deux tableaux donnant les statistiques de temps pour la production des résultats nécessaires à la création du premier article de Statistique Canada et des premiers tableaux CANSIM suivent. Deux ordinateurs ont été utilisés pour la production des résultats. Le premier avait un processeur Intel 4 à 2.4GHz et 524 Mb de RAM tandis que le second avait un processeur Intel 4 à 3.2 GHz et 524 Mb de RAM.

Pour les estimations propres à l'article, peu de répliques bootstrap ont échoué. On a donc opté pour la troisième approche : on a ignoré les répliques bootstrap ayant échoué dans le calcul de la variabilité d'échantillonnage. La contribution des répliques ignorées à la variance échantillonnale était en effet négligeable.

Tableau 2 Temps de calcul requis pour la production des résultats du premier article de Statistique Canada

Variable étudiée	Variable définissant les domaines	Nombre de domaines	Ordinateur	Temps (en minutes)	Nombre de répliques rejetées sur 500
Proportion de l'énergie provenant des lipides	DHHDDDRI	13	1	194	0
	INCOME	8	1	156	0
	AGE1	2	1	148	1
	AGE_REG	10	1	188	0
Proportion de l'énergie provenant des glucides	DHHDDDRI	13	2	162	0
	INCOME	8	2	138	0
	AGE1	2	2	147	1
	AGE_REG	10	2	159	0
Proportion de l'énergie provenant des protéines	DHHDDDRI	13	2	157	0
Produits céréaliers	DHHDDDRI	13	1	207	0
	INCOME	8	1	160	0
	AGE1	2	1	221	0
	AGE_REG	10	1	178	0
Produits laitiers	DIG2	11	2	141	0
	INC_LAIT	12	2	124	8
	AGE2	3	2	114	0
	AGE2_REG	15	2	155	6
Légumes et fruits	DHHDDDRI	13	1	213	0
	INCOME	8	1	160	0
	AGE1	2	1	218	0
	AGE_REG	10	1	188	0
Viandes et substituts	DHHDDDRI	13	2	149	12
	INCOME	8	2	112	13
	AGE1	2	2	148	0
	AGE_REG	10	2	152	0

Les premiers tableaux CANSIM ciblent 130 domaines d'analyse : les 13 plus vieux groupes d'âge et de sexe définis par la variable DHHDDDRI croisés avec les 10 provinces. L'approche des paramètres forcés réplique par réplique (la méthode 5) a été utilisée pour les tableaux CANSIM. Le domaine donneur était le niveau de DHHDDDRI national correspondant au domaine receveur (chaque DHHDDDRI national avait 10 domaines receveurs). Pour la production de ces tableaux, on a donc fait l'hypothèse que les paramètres de l'erreur de mesure nationaux étaient égaux aux paramètres provinciaux. Étant donné que les estimations bootstrap nationales avaient déjà été produites pour les DHHDDDRI, les temps d'exécution supplémentaires dus aux estimations des paramètres du domaine donneur n'étaient pas un facteur dans le choix de la méthode à adopter. Pour calculer le temps total nécessaire pour la production des estimations des tableaux CANSIM, il faut additionner le temps d'exécution du domaine donneur et du domaine receveur. Par exemple, pour l'estimation des pourcentages de l'énergie provenant des lipides des 130 DIG_PRV il a fallu 890+194 minutes (18 heures et 4 minutes).

Tableau 3 Temps de calcul requis pour la production des résultats des premiers tableaux CANSIM

Variable étudiée	Variable définissant les domaines	Nombre de domaines	Ordinateur	Temps (en minutes) nécessaire à la production des estimations du domaine receveur	Nombre de domaines qui n'avaient pas d'estimation avec le poids principal avant l'application forcée des paramètres à cause d'une erreur de type 65
Pourcentage de l'énergie provenant des lipides	DIG_PRV	130	1	890	13
Pourcentage de l'énergie provenant des glucides	DIG_PRV	130	2	874	13
Pourcentage de l'énergie provenant des protéines	DIG_PRV	130	2	600	11

ANNEXE

Description détaillée des solutions aux erreurs découlant de SIDE lors de l'estimation de la variabilité d'échantillonnage à l'aide de répliques Bootstrap

1. Première approche : modifier le domaine d'analyse

On modifiera le domaine d'analyse soit en le divisant en sous-populations (si on suspecte que la distribution de la population totale respecte difficilement les hypothèses du modèle), soit en l'élargissant quand il y aura trop peu de données pour obtenir une bonne estimation.

Élargir le domaine a en général l'avantage de faire diminuer la probabilité d'avoir une estimation de variance négative car le nombre de premier et deuxième rappels se trouvent augmentés. Le principal désavantage est qu'on perd de la précision parce que le domaine final est plus vaste que le domaine qu'on voulait étudier initialement. Un autre désavantage est qu'un domaine plus grand demande plus de temps de calcul.

2. Deuxième approche : modifier les paramètres et les options en entrée

Les paramètres dont les valeurs sont modifiables sont **LINFRAC**, **MAXJP**, **MAXROOT** et **MEANFRAC**. On peut également essayer de faire des changements au niveau du jeu de données CLASSVAR en ajoutant/supprimant les variables indiquant le jour du rappel ou l'indicateur de semaine/fin de semaine du rappel. Parmi ces possibilités, augmenter la valeur de **LINFRAC** est celle qui a le plus de chances de donner des résultats.

3. Troisième approche : ne rien faire

Cette approche est la plus simple, mais on rappelle qu'il faut justifier son utilisation. Les utilisateurs doivent s'assurer qu'il est statistiquement correct d'ignorer les répliques qui échouent et qu'aucun biais n'est introduit. Il faudra aussi vérifier que la variance échantillonnale aura convergé vers une valeur et qu'elle ne fluctuerait pas si d'autres répliques étaient ajoutées. La méthode consiste à calculer la variance échantillonnale avec les répliques qui n'ont pas de message d'erreur fatal. Dans l'exemple, on calculerait la variance entre les 499 répliques (toutes les répliques sauf la 75). On obtiendrait alors comme intervalle de confiance [2041.05, 2578.23] pour la médiane.

4. Quatrième approche : modifier l'estimateur de variance inter-individuelle pour qu'il prenne des valeurs positives ou nulles et utiliser la méthode de Woodruff pour les estimations de percentiles

L'approche consiste à prendre comme estimateur alternatif le maximum entre 0 et l'estimateur du modèle de SIDE. Une variance inter-individuelle nulle revient à dire que tous les individus du domaine étudié ont le même apport nutritionnel habituel. Autrement dit, la distribution de l'apport nutritionnel habituel estimée sera discrète avec toute la masse de probabilité concentrée en la moyenne des apports quotidiens du premier rappel.

Chacun des trois types d'estimation possible à partir de la distribution de l'apport habituel estimée (moyenne, probabilité et percentile) requiert une attention particulière avec le nouvel estimateur de variance inter-individuelle proposé.

1) Estimation De La Moyenne Des Apports Habituels

La moyenne des apports habituels et sa variance échantillonnale peuvent être estimée sans SIDE (en utilisant BOOTVAR par exemple) en calculant simplement la moyenne des apports quotidiens des premiers rappels des répondants. Ce type d'estimation n'est donc pas affecté par les erreurs de SIDE.

2) Estimation Du Pourcentage De La Population Ayant Un Apport Habituel Sous Un Certain Seuil

Pour ce type d'estimation, il faut remarquer que lorsque la variance inter-individuelle est nulle, soit 0% ou 100% de la population est sous le seuil. En effet, avoir une variance entre les apports habituels nulle signifie que tous les individus ont exactement la même consommation en termes d'apport habituel. Par exemple, si on veut estimer la distribution des apports habituels en énergie d'une population, qu'avec le poids principal on a un message d'erreur 65 et que la moyenne des apports quotidiens des premiers rappels est 2000 calories, alors la nouvelle estimation de la variance inter-individuelle est 0 (le maximum entre 0 et la valeur négative que SIDE donnerait) et on conclut que toute la population a un apport habituel de 2000 calories. Ainsi, si on était intéressé au départ à savoir le pourcentage de la population qui a un apport habituel inférieur à 3000 calories, on estimerait ce pourcentage à 100% car 2000 est inférieur à 3000.

Quand on estime l'erreur échantillonnale du pourcentage de la population des hommes de l'Île-du-Prince-Édouard âgés de 31 à 50 ans qui ont un apport habituel en énergie inférieur à 2000 calories (avec ADMDDD comme variable nuisible), on obtient une estimation de variance inter-individuelle négative avec la réplique 75 (donc un message d'erreur 65). La nouvelle estimation de la variance inter-individuelle est 0 pour la réplique 75 (le maximum de la valeur donnée par SIDE qui est négative et 0). La moyenne des premiers rappels est de 2303 calories pour cette réplique. L'estimation bootstrap du pourcentage de la

population sous le seuil de 2000 calories est 0% parce que toute la population a pour apport habituel 2303 calories.

Pour avoir les 500 estimations bootstrap, il s'agit donc de remplacer les estimations bootstrap manquantes par des 0 ou des 1 selon le cas. On remplace donc la valeur manquante pour la réplique 75 par un 0.

3) Estimation D'un Quantile De La Distribution Des Apports Habituels

Pour obtenir 500 estimations bootstrap pour un percentile, il faut combiner la méthode décrite en 2) et la méthode d'estimation par intervalles de confiance de Woodruff. Il faut en résumé : convertir le quantile du nutriment en termes de probabilités, construire l'intervalle de confiance de ces probabilités en suivant la démarche en 2) et finalement revenir dans l'échelle du nutriment.

Voici les étapes à suivre pour obtenir une estimation d'intervalle de confiance pour un percentile de la distribution des apports habituels :

A. Déterminer les quantiles d'intérêt;
B. Calculer les estimations ponctuelles des percentiles déterminés en A avec le poids principal. Si cette étape échoue, la méthode ne fonctionnera pas et il faudra en choisir une autre;
C. Estimer (avec SIDE) la proportion de la population sous les estimations en A pour le poids principal et les 500 poids bootstrap (en utilisant le jeu de données *CUTOFFS*);
D. Remplacer les estimations bootstrap manquantes en C en appliquant la méthode décrite en 2) (pour les estimations de pourcentage de la population sous un certain seuil);
E. Calculer l'intervalle de confiance résultant de C et D;
F. Trouver les quantiles correspondant aux bornes de l'intervalle calculé en E en appliquant l'option *NPTS*=9999 dans le jeu de données *DESC* et en faisant les estimations à partir du poids principal une dernière fois.

Il faudra donc faire 503 estimations à l'aide de SIDE (1 en B, 501 en C combiné avec D et 1 en F) en faisant attention aux paramètres des fichiers d'entrée *DESC* et *CUTOFFS*.

Pour mieux comprendre l'exemple qui suit, il est recommandé de tenter d'obtenir les mêmes estimations avec le logiciel et les données de l'enquête. On veut calculer l'intervalle de confiance de niveau 95% de la médiane de la distribution de l'apport nutritionnel habituel des hommes de l'Île-du-Prince-Édouard âgés de 31 à 50 ans. L'estimation de la médiane, qui est un percentile, nécessitera l'utilisation de la méthode de Woodruff pour la construction de l'intervalle de confiance. Voici chacune des étapes de l'approche en détails :

A. Le quantile d'intérêt est la médiane donc le cinquantième percentile;
B. L'estimation ponctuelle pour ce quantile est 2309.64;

C. Les probabilités d'être sous ce seuil est bien entendu 0.5 pour le poids principal. Pour la première réplique bootstrap on a comme estimation 0.5395; pour la deuxième 0.4278; pour la troisième 0.6575... Il faut noter qu'on n'a pas de valeurs pour la réplique 75 parce qu'on a un message d'erreur 65.
D. La moyenne des premiers rappels de la réplique 75 est 2303. Il faut comparer cette valeur à la valeur obtenue avec le poids principal (soit 2309.64) pour déterminer la valeur bootstrap manquante. On a 1 (ou 100%) pour la réplique 75 car 2303 est plus petit que 2309.64, donc toute la population consomme moins de 2309.64 calories.
E. La variance bootstrap calculée à partir des 500 estimations est 0.0099. L'intervalle de confiance sur les probabilités résultant est [0.3049, 0.6951]. Il reste à convertir cet intervalle de confiance en termes de probabilités dans l'échelle des calories.
F. Une fois qu'on a fixé l'option *NPTS*=9999, on retrouve dans le fichier *PCTU* toutes les estimations nécessaires pour passer de E à F (tous les quantiles de 0.0001 à 0.9999). Les quantiles correspondant aux deux bornes en E sont 2024.93 et 2635.61. L'intervalle de confiance pour la médiane de la distribution des apports nutritionnels moyens des hommes âges de 31 à 50 ans de l'Île-du-Prince-Édouard est donc [2024.93, 2635.61].

IMPORTANT : Une probabilité négative ou plus grande que un n'est pas possible. Pour cette raison, lorsque l'intervalle de confiance en E a des valeurs négatives et/ou plus grandes que 1, il faut ramener l'intervalle de confiance à un intervalle entre 0 et 1. Par exemple, si on obtient en E l'intervalle [-0.01, 1.03], l'intervalle qu'il faudrait utiliser à l'étape F serait [0, 1].

IMPORTANT : Il faut noter qu'obtenir une borne de 0 ou de 1 pour l'intervalle en E signifie que l'on perd l'information sur cette (ou ces) borne(s) à l'étape F. Ceci survient quand la variabilité est très grande dans le domaine étudié, c'est-à-dire que l'enquête ne nous donne pas beaucoup d'information sur ce domaine. Dans l'exemple, si on avait obtenu [0.0000, 0.6951] comme intervalle en E, en F on aurait eu comme intervalle [*x* calories, 2635.61 calories], où *x* est la plus petite valeur d'apport habituel en énergie possible. Autrement dit, on obtient un intervalle de confiance unilatéral alors qu'on en voulait un bilatéral. De façon similaire, si on avait eu [0.3049, 1.0000] à l'étape E, l'intervalle en F aurait été [2024.93 calories, *y* calories], où *y* est la plus grande valeur possible d'apport habituel. Finalement, si l'intervalle en E avait été [0, 1], en F l'intervalle lui aurait été [*x* calories, *y* calories], c'est-à-dire qu'on n'aurait eu aucune information sur la médiane en bout du compte à part l'estimation ponctuelle de 2309.64. Autrement dit, l'estimation ponctuelle est 2309.64, mais au niveau de confiance de 95% on conclut que la médiane pourrait être n'importe quelle valeur. Si on a estimé qu'au moins la moitié des répliques bootstrap allaient avoir un message d'erreur 65, on est à peu près certain d'avoir une borne non informative (et peut-être même les deux).

Rien n'empêche d'appliquer cette méthode lorsque toutes les répliques bootstrap fonctionnent au départ. Les intervalles de confiance obtenus en faisant les estimations de la

manière normale et en appliquant cette approche seront alors différents, mais ils seront tous les deux valides si les hypothèses de base sont respectées.

5. Cinquième approche : forcer la valeur des paramètres relatifs à l'erreur de mesure réplique par réplique

Les valeurs forcées (501 poids fois 2 paramètres) doivent provenir des données de l'enquête. Forcer les paramètres revient à faire l'hypothèse que la variance inter-individuelle dans la population étudiée et dans la population qui fournit les paramètres est la même. Il est plus raisonnable de faire ce genre d'hypothèse avec un macro nutriment (comme l'énergie par exemple) qu'avec un nutriment ou un aliment plus local ou plus rare (par exemple les fruits de mer).

Si on prend encore une fois l'exemple de l'apport nutritionnel habituel en énergie des hommes de 31 à 50 ans de l'Île-du-Prince-Édouard, la réplique 75 donnera une estimation de variance inter-individuelle négative. Pour appliquer la méthode, on pourrait prendre comme domaine donneur les hommes de 31 à 50 ans du Canada.

Les valeurs des paramètres à forcer (variance de l'erreur de mesure et quatrième moment de l'erreur de mesure) seront alors 0.5903 et 4.1645 pour le poids principal; 0.5702 et 4.2026 pour la première réplique; 0.6676 et 4.0390 pour la deuxième; 0.6133 et 4.3896 pour la troisième et ainsi de suite. La valeur de l'estimation ponctuelle de la médiane est alors 2324.21. L'intervalle de confiance estimé est [2052.62, 2595.80].

On remarque que dans les exemples, les méthodes 3, 4 et 5 donnent des intervalles comparables. Ne rien faire a peu d'impact dans l'exemple car il n'y a qu'une réplique qui est en jeu. Quand le nombre de répliques est plus grand, la différence entre les intervalles de confiance est aussi plus grande. Il faut alors être plus prudent dans le choix de la méthode qu'on adoptera.

6. Sixième approche : forcer la valeur des paramètres relatifs à l'erreur de mesure avec les mêmes valeurs pour toutes les répliques

Les valeurs forcées peuvent provenir des données de l'enquête comme elles peuvent provenir d'une autre enquête (ou des connaissances d'un expert à la limite). L'approche consiste à remplacer les valeurs de variance et du quatrième moment pour le poids principal et toutes les répliques (incluant celles qui n'ont pas échoué), mais avec les mêmes deux valeurs à chaque fois.

Si on prend une dernière fois la population des hommes de 31 à 50 ans de l'Île-du-Prince-Édouard et qu'on cherche à estimer la médiane des apports habituels en énergie, on aura un échec avec la réplique 75. Si le domaine donneur est les hommes de 31 à 50 ans du Canada, on forcera les valeurs des paramètres obtenus avec ce domaine et le poids principal. Ces valeurs sont 0.5903 et 4.1645 et elles seront forcées pour le poids principal et chacune des répliques du domaine receveur. La valeur de l'estimation ponctuelle de la médiane est alors 2324.21 et l'intervalle de confiance estimé est [2224.83, 2423.59]. On note que l'intervalle

produit est environ deux fois plus petit que l'intervalle obtenu avec la méthode 4 (où on force des valeurs différentes à chaque répliques) parce qu'on a ignoré la variabilité échantillonnale due à l'estimation des deux paramètres forcés. En effet, forcer des valeurs différentes à chaque réplique est équivalent à tenir compte de cette variabilité.