Guide du fichier de microdonnées à grand diffusion

Série d'enquêtes sur les perspectives canadiennes 4 : Les sources d'information consultées pendant la pandémie

Le document suivant est un complément du guide de l'utilisateur de la série de l'enquête sur les perspectives canadiennes : SEPC4– Les sources d'information consultées pendant la pandémie. Le document présente les différences entre le fichier maître et le fichier de microdonnées à grande diffusion (FMGD) en plus de fournir un court guide pour analyser, publier et interpréter de manière appropriée les données fournies dans le FMGD. Pour avoir plus de détails concernant la SEPC, incluant la population cible, le plan d'échantillonnage ou le plan de pondération, veuillez vous référer au guide de l'utilisateur *SEPC4 Guide analytique*.

1. Documentation

Le questionnaire de la SEPC4 – Les sources d'information consultées pendant la pandémie sert de référence au fichier de données FMGD de la SEPC4. Le questionnaire affichera des questions additionnelles qui ne sont pas présentées sur le FMGD en raison de suppressions et de méthodes de restrictions qui ont été appliquées aux données.

2. Conversion du fichier maître en fichier de microdonnées à grande diffusion

L'approche utilisée pour créer le FMGD vise à équilibrer les exigences pour maintenir la confidentialité des répondants en minimisant les risques de divulgation tout en fournissant les données les plus utiles aux utilisateurs. Certaines différences entre le fichier maître et le FMGD de la SEPC4 incluent le retrait de la province du fichier et de certaines variables reliées au travail (voir l'Annexe A). Pour d'autres variables, des catégories de réponse ont été regroupées (voir l'Annexe B). De plus, les poids bootstrap (utilisés dans le calcul des variances) ne sont pas fournis avec le FMGD. Les utilisateurs devraient appliquer les facteurs fournis au Tableau 1 de la Section 3. Ces facteurs vont donner une approximation de la précision des estimations. **Lorsque des différences entre les valeurs du fichier maître et celles du FMGD sont notables, les valeurs du fichier maître sont considérées comme la source ayant préséance.**

3. Lignes directrices pour les analyses statistiques

Poids d'enquête: La SEPC4 est basée sur un plan d'échantillonnage complexe incluant de la stratification, plusieurs degrés de sélection et des probabilités inégales de sélection des répondants. Le plan d'échantillonnage n'était pas auto-pondéré. Lors de la production d'estimations simples, incluant la production de tableaux statistiques, les utilisateurs **doivent** appliquer les poids d'enquête appropriés. Si les poids appropriés ne sont pas utilisés, les estimations produites à partir du FMGD ne pourront pas être considérées comme étant représentatives de la population enquêtée.

Estimation de la variance pour le FMGD: Les variances produites par les *progiciels* statistiques s'appuient généralement sur des formules simples qui ne prennent pas en considération la complexité des plans d'échantillonnage. Ces formules risquent de sous-estimer la vraie variance des estimations de la SEPC4. Par conséquent, l'enquête a utilisé une méthode de ré-échantillonnage appelée le bootstrap. Pour des raisons de confidentialité, les poids bootstrap ne sont pas fournis avec le FMGD, mais ils ont été utilisés pour créer des facteurs d'ajustement qui peuvent être appliqués aux estimations de la variance, aux erreurs types et aux coefficients de variation calculés par un *progiciel* statistique. Les facteurs d'ajustement du Tableau 1 devraient être multipliés par la variance/l'erreur type/le coefficient de variation, produits par un logiciel statistique, afin de tenir compte du plan complexe de l'enquête.

Tableau 1. Facteurs d'ajustement à appliquer aux mesures de précision

Catégorie d'âge	Facteur d'ajustement (aussi appelé effet de plan) pour une **Variance**	Facteur d'ajustement pour l'**Erreur type**
Plus d'un groupe d'âge combiné	3.5	1.8708
15-24	2.5	1.5811
25-34	2.5	1.5811
35-44	2.5	1.5811
45-54	2.5	1.5811
55-64	2.5	1.5811
65-74	3.0	1.7321
75+	3.0	1.7321

L'exemple suivant a été tiré de SEPC4. Supposons, par exemple, que nous voulions connaître la proportion des personnes qui ont dit avoir une excellente ou une très bonne santé mentale. La première étape consisterait à créer un fichier en ne conservant que les observations pour lesquelles il y a une réponse à la question MH_30 et de créer une variable indicatrice :

$$\text{santé_mentale} = \begin{cases} 1 & \text{si MH_30} \in \{1,2\} \text{ (c.-à-d. excellente ou très bonne santé mentale)} \\ 0 & \text{si MH_30} \in \{3,4,5\} \text{ (c.-à-d. bonne, passable ou mauvaise santé mentale)} \end{cases}$$

Supposons que ce fichier s'appelle fmgd_sante. Pour calculer la proportion moyenne des personnes ayant une très bonne ou excellente santé mentale, pour tous les âges regroupés, une procédure SAS *proc means* avec les poids peut être utilisée.

Exemple de code SAS :

```
proc means data=fmgd_sante mean stderr;
      var sante_mentale;
      weight PERS_WGT;
run;
```

De cette procédure, nous obtenons que 0.5534 de la population a une très bonne ou excellente santé mentale (équivalent à dire que 55% de la population a une très bonne ou excellente santé mentale).

L'erreur type (S.E.) obtenue pour cette estimation par la procédure précédente est 0. 0076601. Cependant, le plan de sondage n'est pas pris en compte dans cette mesure et sous-estime l'erreur type réelle.

Pour estimer l'erreur type adéquate, le facteur d'ajustement pour l'erreur type mentionné dans le tableau 1 doit être appliqué :

$$\text{S. E.} = \underbrace{0.0076601}_{\text{obtenu par la procedure ci-dessus}} \cdot \underbrace{1.8708}_{\text{Tableau 1}}$$

$$= 0.0143$$

Une fois que l'erreur type S.E. est calculée, d'autres mesures telles que le coefficient de variation (CV) et l'intervalle de confiance (I.C.) peuvent être calculés.

Le coefficient de variation est calculé de la façon suivante :

$$\text{CV} = \frac{\text{S.E.}}{\text{moyenne}} = \frac{0.0143}{0.5534} = 0.02584$$

Pour obtenir l'intervalle de confiance à 95%:

$$\begin{aligned}\text{I.C.} &= [p - 1.96(\text{S.E.}), p + 1.96(\text{S.E.})] \\ &= [0.5534 - 1.96(0.0143), 0.5534 + 1.96(0.0143)] \\ &= [0.5254, 0.5814]\end{aligned}$$

L'intervalle de confiance peut aussi être exprimé sous la forme [52.5%, 58.1%].

Lignes directrices pour l'arrondissement: Il est vivement conseillé aux utilisateurs de respecter les lignes directrices qui suivent en ce qui concerne l'arrondissement de telles estimations :

a) Les estimations dans le corps principal d'un tableau statistique doivent être arrondies à la centaine près à l'aide de la technique d'arrondissement classique. Selon cette technique, si le premier ou le seul chiffre à supprimer se situe entre 0 et 4, le dernier chiffre à conserver ne change pas. Si le premier ou le seul chiffre à supprimer se situe entre 5 et 9, le dernier chiffre à conserver est augmenté de 1. Par exemple, selon la technique d'arrondissement classique à la centaine près, si les deux derniers chiffres se situent entre 00 et 49, ils sont remplacés par 00 et le chiffre précédent (le chiffre des centaines) reste inchangé. Si les derniers chiffres se situent entre 50 et 99, ils sont remplacés par 00 et le chiffre précédent est augmenté de 1.

b) Les sous-totaux et totaux marginaux des tableaux statistiques doivent être calculés à partir de leurs composantes non arrondies correspondantes, puis être arrondis à leur tour à la centaine près à l'aide de la technique d'arrondissement classique.

c) Les moyennes, les taux et les pourcentages doivent être calculés à partir de composantes (c'est-à-dire des numérateurs ou des dénominateurs) non arrondies, puis être arrondis à leur tour à une décimale à l'aide de la technique d'arrondissement classique. Dans le cas d'un arrondissement classique à un seul chiffre, si le dernier ou le seul chiffre à supprimer se situe entre 0 et 4, le dernier chiffre à conserver ne change pas. Si le premier ou le seul chiffre à supprimer se situe entre 5 et 9, le dernier chiffre à conserver est augmenté de 1. Les proportions et rapports doivent d'abord être calculés à partir de composantes non arrondies, puis être arrondis à leur tour à la troisième décimale à l'aide de la technique d'arrondissement classique.

d) Les sommes et les différences d'agrégats doivent être calculées à partir de leurs composantes non arrondies correspondantes, puis être arrondies à leur tour à la centaine près (ou à la décimale près). Les sommes et les différences de pourcentages (ou rapports) doivent être calculées à partir de leurs composantes non arrondies correspondantes, puis être arrondies à leur tour à la décimale près (ou à la troisième décimale) à l'aide de la technique d'arrondissement classique.

e) Dans les cas où, en raison de limitations d'ordre technique ou de toute autre limite, une technique d'arrondissement autre que la technique classique est utilisée, produisant des estimations à être publiées ou autrement diffusées différentes des estimations correspondantes publiées par Statistique Canada, il est vivement conseillé aux utilisateurs d'indiquer la raison de ces différences dans le ou les documents de la publication .

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

Directives de qualité pour les estimations: Avant de diffuser et/ou de publier toute estimation, les utilisateurs devraient considérer le niveau de qualité des estimations. La section suivante traitera de la qualité en termes d'erreur d'échantillonnage. Il existe différentes méthodes pour mesurer et rapporter l'erreur d'échantillonnage. Il est considéré comme une pratique exemplaire à Statistique Canada de rapporter l'erreur d'échantillonnage d'une estimation selon son intervalle de confiance à 95%. L'intervalle de confiance devrait être diffusé avec l'estimation, et ce dans le même tableau que l'estimation. En plus des intervalles de confiance, les estimations sont catégorisées dans l'une des trois catégories de qualité suivantes :

> **Catégorie A:** Les estimations peuvent être diffusées sans qu'une mise en garde ne soit nécessaire. Les utilisateurs de données devraient utiliser les intervalles de confiance à 95% pour décider si la qualité des estimations est suffisante.

Catégorie E – Qualité marginale: Les estimations et les intervalles de confiance sont jugés comme étant de qualité marginale. Les estimations et les intervalles de confiance devraient être identifiés par la lettre E (ou un identifiant similaire) et devraient être accompagnés d'une mise en garde aux utilisateurs d'utiliser l'estimation avec précaution.

Catégorie F – Qualité insuffisante: Les estimations et les intervalles de confiance sont jugées de qualité insuffisante et il n'est pas recommandé de les publier. Les estimations contiennent un haut niveau d'instabilité, les rendant non fiables et potentiellement trompeuses. Si l'utilisateur choisit malgré tout de publier des estimations de qualité insuffisante, malgré les avertissements à propos de leur précision défaillante, les estimations devraient être accompagnées d'un avertissement. L'utilisateur devrait reconnaître les avertissements qui ont été donnés et entreprendre de ne pas diffuser, présenter ou rapporter les estimations, directement ou indirectement, sans cette mise en garde. Les estimations devraient être identifiées par la lettre F (ou un identifiant similaire) et l'avertissement suivant devrait accompagner les estimations et les intervalles de confiance:

« Soyez avisés que ces estimations et intervalles de confiance [identifiés par la lettre F] ne répondent pas aux normes de qualité de Statistique Canada. Les conclusions tirées de ces données ne sauraient être fiables et seront fort probablement erronées. »

Tableau 2. Taille échantillonnale exigée par catégorie de qualité

Type de statistique	Catégorie F	Catégorie E	Catégorie A
Proportion	$n < 90$	$90 \leq n \leq 180$	$n > 180$
Compte pondéré	$m < 90$	$90 \leq m \leq 180$	$m > 180$

Notation utilisée dans le Tableau 2:

n: Taille échantillonnale du domaine. Pour les proportions, n est le compte non-pondéré du nombre de répondants du dénominateur de la proportion.

m: Compte non-pondéré du nombre de répondants ayant des valeurs non-nulles contribuant à l'estimation.

5. Notes supplémentaires pour l'analyse

Variables sur le travail

PEMPSTC

- Le concept d'employé et absent du travail dans la SEPC4 n'est pas équivalent aux concepts de l'Enquête sur la population active (EPA) et aucune comparaison directe ne devrait être faite. Dans la SEPC4, «employé» (catégories 1 à 3 de la variable PEMPSTC) inclut toutes les personnes qui ont travaillé ou qui se sont absentés pour quelles que raisons que ce soit, incluant les mises-à-pied temporaires. Celles qui ont été temporairement mis à pied ne sont pas traitées comme étant absentes ou employées dans l'EPA.
- Étant donné la rapidité avec laquelle la situation évoluait au moment de l'enquête (c.-à-d. les directives de la santé publique, les ordres du gouvernement), la semaine de référence pour les questions sur l'état d'employé (12 au 18 juillet 2020) devrait être clairement mentionnée lorsque des résultats sont rapportés. Les informations appropriées devraient être fournies au sujet de la nature et de l'ampleur des restrictions en place à ce moment.

Annexe A

Le tableau suivant contient la liste des variables qui ont été retirées durant la création du FMGD.

Variables retirées

Variable	Description
Province	Province de résidence
Region	Région de résidence
CSizeMiz	Taille communauté et zones d'influence métropolitaine
LM_05	Au cours de cette semaine, avez-vous travaillé à un emploi ou une entreprise?
LM_10	Au cours de cette semaine, aviez-vous un emploi ou une entreprise dont vous vous êtes absenté(e)?
LM_15	Quelle était la raison principale pour laquelle vous vous êtes absenté(e) du travail cette semaine?
TOTHHSIZ	Taille totale du ménage

Annexe B

Variables recodées ou plafonnées

Description de la variable	Étiquette recodée	
Âge du répondant (originale: Curr_Age) (recodée: AgeGrp)	1 2 3 4 5 6 7	15 à 24 ans 25 à 34 ans 35 à 44 ans 45 à 54 ans 55 à 64 ans 65 à 74 ans 75 ans et plus
Taille du ménage (originale: HHLDSIZE) (recodée: HHLDSIZC)	1 2 3 4 5	1 2 3 4 5 et plus
Niveau d'éducation (originale: EDUC_LVL) (recodée: PEDUC_LC)	1 2 3 4 5 6 7	Sans diplôme d'études secondaires ou équivalent Diplôme études secondaires/certificat, équivalence études secondaires Certificat ou diplôme d'une école de métiers Certificat/diplôme Collège/Cégep/Autre établissement non universitaire Certificat/diplôme universitaire au-dessous du niveau du bac Baccalauréat (p. ex. B.A., B.Sc., LL.B.) Certificat/diplôme/grade univers. au-dessus du niveau du bac
Type de logement (originale: DWELCODE) (recodée: DWELCODC)	1 2 3 4	Maison individuelle Immeuble d'appartements moins de 5 étages Immeuble d'appartements 5 étages ou plus Autre
Situation familiale (originale: MARSTAT) (recodée: MARSTATC)	1 2 3 4	Marié En union libre Veuf/Séparé/Divorcé Célibataire/Jamais marié
Statut d'immigration (originale: IMMIGRNT) (recodée: IMMIGRNC)	1 2	Né au Canada Résident permanent et résident non permanent
Statut d'emploi (originale: EMPST) (recodée: PEMPSTC)	1 2 3 4	Employé et au travail au moins une partie semaine référence Employé et absent travail pour raisons non reliées au COVID-19 Employé et absent du travail dû à la COVID-19 Sans emploi

	9	Non déclaré
Principale source d'information pour en savoir plus sur la COVID-19 (original: BH_05) (recoded: BH_05C)	1 2 3 4 5 6 7 8 9 10 11 12 99	Médias d'informations, y compris des sources locales, nationales et internationales Organisme de santé fédéral Organisme de santé provincial ou territorial Organisme de santé municipal Déclarations quotidiennes faites par des responsables de la santé publique et des dirigeants politiques fédéraux Déclarations quotidiennes faites par des responsables de la santé publique et des dirigeants politiques provinciaux ou territoriaux Médias sociaux Famille, amis ou collègues Professionnels de la santé Lieu de travail Autre Je ne cherche pas d'information à ce sujet Non déclaré
Raison l'exactitude non validée – Ne savais pas comment vérifier/trop difficile (original: FC_20C, FC_20E) (recoded: FC_20CE)	1 2 6 9	Raison l'exactitude non validée – Ne savais pas comment vérifier Raison l'exactitude non validée – Savais comment vérifier Enchaînement valide Non déclaré