9 LIGNES DIRECTRICES POUR LA DIFFUSION (QUALITÉ DES DONNÉES

Les utilisateurs de microdonnées doivent appliquer les règles d'évaluation de la qualité des données présentées ci-dessous à toutes les estimations qu'ils produisent; ils ne doivent retenir que les estimations qui répondent aux critères de diffusion. Les estimations qui ne satisfont pas aux conditions de diffusion ne sont pas fiables.

Introduction

Les lignes directrices relatives à la diffusion et à la publication s'appuient sur le concept de la *variabilité d'échantillonnage* qui permet de déterminer si les estimations obtenues à partir des fichiers de microdonnées sont fiables. La variabilité d'échantillonnage est l'erreur qui, dans les estimations, est attribuable au fait que l'enquête porte sur un échantillon plutôt que sur l'ensemble de la population. Le concept d'*erreur-type* et les concepts connexes de *coefficient de variation* et d'*intervalle de confiance* donnent une indication de l'importance de la variabilité d'échantillonnage.

L'erreur-type et le coefficient de variation ne permettent pas de mesurer les biais systématiques des données de l'enquête qui pourraient influer sur les estimations. Ils sont plutôt fondés sur l'hypothèse selon laquelle les erreurs d'échantillonnage suivent une courbe normale.

Suivant cette hypothèse, il est possible d'estimer dans quelle mesure différents échantillons prélevés à partir d'un même plan de sondage et comportant le même nombre d'observations donneraient lieu à des résultats divergents. Cette mesure indique la marge d'erreur susceptible d'être comprise dans les estimations tirées de l'unique échantillon utilisé. Pour une description détaillée des mesures de la variabilité d'échantillonnage, voir A. Satin et W. Shastry, *L'échantillonnage : un guide non mathématique*, Statistique Canada, nº 12-602F au cat.

Erreurs reliées aux enquêtes

L'enquête permet de produire des estimations fondées sur les renseignements recueillis auprès d'un échantillon de personnes et concernant ces dernières. Les estimations obtenues seraient peut-être sensiblement différentes si on réalisait un recensement exhaustif en reprenant le même questionnaire, les mêmes intervieweurs et superviseurs, les mêmes méthodes de traitement, etc. que pour l'enquête elle-même. L'écart entre les estimations découlant de l'échantillon et celles que donnerait un dénombrement exhaustif réalisé dans des conditions comparables est appelé erreur d'échantillonnage de l'estimation.

Des erreurs qui ne sont pas reliées à l'échantillonnage peuvent se produire à presque toutes les étapes des opérations d'enquête. Les intervieweurs peuvent avoir mal compris les instructions, les personnes interrogées peuvent se tromper en

répondant aux questions, les réponses peuvent être mal entrées sur l'ordinateur, et des erreurs peuvent être faites au moment du traitement et de la totalisation des données. Il s'agit là d'autant d'erreurs non dues à l'échantillonnage.

Lorsque le nombre d'observations est élevé, les erreurs aléatoires ont peu d'effet sur les estimations calculées à partir des résultats de l'enquête. Toutefois, les erreurs systématiques contribuent à biaiser les estimations. Un temps et des efforts considérables sont consacrés à la réduction des erreurs non dues à l'échantillonnage dans le cadre de l'enquête. À chacune des étapes du cycle de collecte et de traitement des données, on applique des mesures d'assurance de la qualité pour contrôler la qualité des données. Au nombre de ces mesures figurent le recours à des intervieweurs hautement qualifiés, une formation poussée des intervieweurs au chapitre des procédures et du questionnaire de l'enquête, l'observation des intervieweurs en vue de cerner les problèmes liés à la conception du questionnaire ou à une mauvaise compréhension des instructions, des contrôles visant à réduire au minimum les erreurs de saisie des données ainsi que des vérifications du codage et des contrôles de la qualité ayant pour but d'assurer la logique du traitement.

L'incidence de la non-réponse sur les résultats d'enquête constitue une source importante d'erreurs non dues à l'échantillonnage. La non-réponse peut être partielle (le fait de ne pas répondre à une ou à

quelques questions) ou totale. Les cas de non-réponse totale se produisent lorsque l'intervieweur ne peut joindre le répondant, lorsqu'aucun membre du ménage n'est en mesure de fournir les renseignements ou lorsque le répondant refuse de participer à l'enquête. On traite la non-réponse totale en redressant le poids des ménages qui prennent part à l'enquête de façon à compenser pour les ménages qui n'y participent pas (consulter le chapitre 6).

Dans la plupart des cas, la non-réponse partielle à l'enquête se produit lorsque le répondant ne comprend pas ou interprète mal une question, lorsqu'il refuse de répondre à une question, lorsqu'il ne peut se rappeler les renseignements requis ou lorsqu'il ne peut fournir des renseignements par procuration.

Comme il est impossible de soustraire à l'erreur d'échantillonnage les estimations découlant d'une enquête par échantillon, les chercheurs, soucieux d'appliquer de saines pratiques sur le plan statistique, fournissent aux utilisateurs certaines indications quant à l'importance de l'erreur d'échantillonnage. La présente section de la documentation donne un aperçu des mesures de l'erreur d'échantillonnage que Statistique Canada utilise fréquemment; le Bureau recommande vivement aux utilisateurs d'en tenir compte au moment de produire des estimations à partir des fichiers de microdonnées.

L'erreur-type des estimations découlant des résultats d'enquête constitue le fondement de la mesure de la taille potentielle des erreurs d'échantillonnage.

Tests d'hypothèse intégrés aux progiciels

L'Enquête sur les voyages des Canadiens est fondée sur un plan d'échantillonnage complexe comportant une stratification et de multiples degrés de sélection ainsi que des probabilités inégales de sélection des répondants. L'utilisation des données provenant d'enquêtes aussi complexes présente des difficultés aux analystes dans la mesure où le plan d'enquête et les probabilités de sélection influent sur les méthodes d'estimation et de calcul de la variance qui doivent être adoptées.

Bien que de nombreuses méthodes d'analyse qui font partie des progiciels statistiques permettent d'utiliser des poids, la définition ou la signification du poids adoptée dans ces procédures diffère de celle qui convient à une enquête par sondage, de sorte que si les estimations faites au moyen de ces progiciels sont exactes dans bien des cas, les variances calculées n'ont pratiquement aucune signification.

Dans le cas de nombreuses techniques d'analyse (par exemple, la régression linéaire, la régression logistique et l'analyse de variance), il existe un moyen de rendre l'application des progiciels standards plus significative. Si l'on procède à un rééchelonnement des poids des enregistrements de manière à obtenir un poids moyen de un (1), les

résultats obtenus à l'aide de ces logiciels standards seront plus raisonnables. Ils tiendront compte des probabilités inégales de sélection mais non de la structure stratifiée et de la répartition en grappes du plan de sondage. On peut effectuer ce rééchelonnement en divisant chaque poids par le poids moyen global avant d'entreprendre l'analyse.

Afin de donner aux utilisateurs le moyen d'évaluer la qualité des estimations totalisées, Statistique Canada a produit un ensemble de tableaux de degrés approximatifs de variabilité de l'échantillonnage (généralement désignés par le terme *tableaux de CV*) pour l'Enquête sur les voyages des Canadiens. Ces tableaux peuvent servir à obtenir les coefficients de variation approximatifs des estimations de type nominal et des proportions. Consulter le chapitre 10 pour plus de détails.

Effectif minimum des groupes visés par les estimations aux fins de la diffusion

Avant de diffuser ou de publier des estimations tirées de ces fichiers de microdonnées, les utilisateurs doivent commencer par établir le nombre de répondants visés par le calcul de l'estimation. Si ce nombre est inférieur à 30, l'estimation pondérée ne peut être diffusée, peu importe la valeur de son coefficient de variation. Dans le cas d'estimations pondérées fondées sur des échantillons comportant 30 unités ou plus, les utilisateurs doivent déterminer le coefficient de variation de l'estimation arrondie et suivre les lignes directrices présentées au tableau 5.

Lorsque l'estimation non pondérée est satisfaisante, l'utilisateur doit vérifier si l'estimation pondérée répond aux critères de diffusion. Les seuils relatifs à la diffusion des estimations pondérées tirées de l'EVC apparaissent au chapitre 10 qui porte sur la qualité des données.

Utilisation du coefficient de variation (CV)

L'erreur-type d'une estimation est souvent exprimée sous la forme d'un pourcentage de l'estimation elle-même; dans ce cas, elle est appelée *coefficient de variation*. Si l'erreur-type est mesurée selon les mêmes unités que l'estimation, le coefficient de variation, quant à lui, est simplement un rapport, ce qui facilite son utilisation à titre de critère de fiabilité des estimations.

Supposons par exemple que, à partir des résultats de l'enquête, on estime à 25,9 % la proportion des Canadiens âgés de 15 ans et plus ayant effectué au

TABLEAU 5. Niveaux acceptables du coefficient de variation

Coefficient de variation approximatif (en %)	**Restrictions relatives à la diffusion**
0.0 - 16.5	ACCEPTABLE. On peut envisager une diffusion générale sans restriction des estimations.
16.6 - 25.0	MÉDIOCRE. On peut envisager une diffusion générale sans restriction des estimations si l'on joint une mise en garde aux utilisateurs quant à la variabilité élevée d'échantillonnage liée aux estimations.
25.1 - 33.3	CONFIDENTIEL. On ne peut envisager une diffusion générale sans restriction des estimations que si l'on obtient, contre recouvrement des coûts, les coefficients exacts de variation et que ceux-ci se révèlent acceptables, à défaut de quoi ces estimations ne devraient être ni utilisées ni diffusées.
33.4 ou plus	INACCEPTABLE. En aucun cas les estimations ne doivent être utilisées ou diffusées.

moins un voyage en mars 1996 et que l'erreur-type de cette estimation est établie à 0.009. Dans ce cas, le coefficient de variation de cette estimation se calcule comme suit :

$$\left(\frac{.009}{.259}\right) \; x \; 100\% \; = \; 3.47\%$$

Les coefficients de variation (CV) sont calculés au moyen de la formule de variance applicable à l'échantillonnage aléatoire simple et comportent un facteur qui tient compte du fait qu'il s'agit d'un échantillonnage en grappes à plusieurs degrés. On détermine ce facteur, désigné par le terme *effet du plan de sondage*, en calculant d'abord les effets du plan de sondage pour une large gamme de caractéristiques et en choisissant ensuite parmi celles-ci une valeur prudente qui sera utilisée dans les tables de recherche et qui s'appliquera à l'ensemble des caractéristiques.

L'utilisateur est invité à se reporter au Rapport sur la qualité des données (chapitre 10) pour obtenir de plus amples renseignements sur les effets du plan de sondage, les tailles d'échantillon et les chiffres de population selon la province, utilisés pour produire les tableaux des degrés approximatifs de variabilité de l'échantillonnage.

Il convient de noter que des tableaux des degrés approximatifs de variabilité de l'échantillonnage sont également disponibles pour les poids du

voyage-personne, voyage-ménage, nuitée-personne et voyage-dépense.

Dans le Rapport sur la qualité des données figurant dans le présent guide, un ensemble de tableaux des degrés approximatifs de variabilité de l'échantillonnage sont fournis pour donner aux utilisateurs de microdonnées certains coefficients de variation approximatifs visant des ensembles particuliers d'estimations, par exemple, l'ensemble des estimations relatives à une province donnée. Dans la plupart des cas, ces coefficients de variation approximatifs seront suffisants pour déterminer s'il convient ou non de diffuser une estimation. Le Rapport sur la qualité des données (chapitre 10) explique comment obtenir le CV approximatif à partir des tableaux, suivant que l'estimation est un simple chiffre de population ou un pourcentage, une différence ou un rapport visant des sous-groupes de la population. Le CV de totaux ou de moyennes numériques est généralement plus élevé que celui du chiffre de la population auquel il se rapporte.

Tous les coefficients de variation apparaissant dans les tableaux des degrés approximatifs de variabilité de l'échantillonnage sont approximatifs et, par conséquent, non officiels. On peut se procurer, contre recouvrement des coûts, les estimations de la variance réelle de variables particulières auprès de Statistique Canada. L'utilisation d'estimations de la variance réelle permet aux utilisateurs de diffuser des estimations qui ne pourraient l'être autrement, c.-à-d. les estimations dont le coefficient de

variation se situe dans la fourchette du niveau «confidentiel».

Note : Si le nombre d'observations visées par une estimation est inférieur à 30, l'estimation pondérée ne doit pas être diffusée, peu importe la valeur de son coefficient de variation. Cette règle est appliquée parce que les formules utilisées pour estimer la variance ne tiennent pas dans le cas d'échantillons de petite taille.