

Enquête sur les véhicules au Canada – Exactitude des données

Malgré tous les efforts que l'on peut faire pour maintenir une norme de qualité élevée tout au long des opérations d'enquête, les estimations qui en résultent sont inévitablement sujettes à un certain degré d'erreur. L'erreur d'enquête totale est définie comme la différence entre l'estimation de l'enquête et la valeur réelle de la population que l'estimation de l'enquête tente d'atteindre. L'erreur d'enquête totale est formée de deux types d'erreur : l'erreur d'échantillonnage et les erreurs non dues à l'échantillonnage.

Erreur d'échantillonnage

Lorsqu'on tire un échantillon d'une population, les estimations fondées sur les données d'échantillon ne sont pas toujours exactement les mêmes que les résultats qu'aurait produits un recensement de la même population. Les deux résultats risquent de diverger, puisque seules les données des unités échantillonnées sont utilisées. Dans le cas d'un recensement, il n'y a pas d'erreur d'échantillonnage.

La différence entre les estimations d'une enquête basée sur un échantillon et d'un recensement réalisé dans les mêmes conditions s'appelle l'erreur d'échantillonnage d'une estimation d'enquête. L'erreur d'échantillonnage dépend de facteurs comme la taille de l'échantillon, le plan d'échantillonnage, la variabilité de la caractéristique observée dans la population et la méthode d'estimation. Si la population est très hétérogène, comme la population des véhicules immatriculés, il faut un grand échantillon pour obtenir des estimations fiables.

L'erreur d'échantillonnage se mesure par une quantité statistique appelée l'erreur-type. Cette quantité reflète la variabilité espérée de l'estimation d'enquête d'une caractéristique donnée de la population si l'on utilise un échantillonnage répété. La vraie valeur de l'erreur-type est, bien sûr, inconnue, mais elle peut être estimée, à partir de l'échantillon. L'erreur-type estimée est utilisée sous forme de mesure relative appelée le coefficient de variation (ou CV). Cette mesure n'est autre chose que l'erreur-type estimée, exprimée en pourcentage de la valeur de l'estimation d'enquête. Par conséquent, plus le CV est petit, plus l'estimation est fiable.

Erreurs non dues à l'échantillonnage

L'erreur d'échantillonnage n'est qu'une composante de l'erreur d'enquête totale. Les autres erreurs découlant de toutes les phases d'une enquête sont des erreurs non dues à l'échantillonnage. À mesure que la taille de l'échantillon se rapproche de la taille de la population, la composante « erreur

d'échantillonnage » de l'erreur d'enquête totale devrait diminuer. Mais il n'en va pas nécessairement de même pour la composante « erreurs non dues à l'échantillonnage ». En effet, ce type d'erreur peut surgir lorsqu'un répondant donne des renseignements erronés ou ne répond pas à certaines questions, lorsqu'une unité de la population d'intérêt est oubliée ou observée plus d'une fois, lorsqu'une unité qui est hors du champ de l'enquête y est incluse par erreur ou lorsque des erreurs surviennent en cours de traitement des données, comme des erreurs de codage et de saisie.

Certaines erreurs non dues à l'échantillonnage s'annulent sur un grand nombre d'observations, mais les erreurs systématiques (c.-à-d. celles qui n'ont pas tendance à s'annuler) contribuent à un biais des estimations. Ainsi, dans le cas de l'EVC, si des personnes qui utilisent leur véhicule plus que la moyenne ont toujours tendance à ne pas répondre à l'enquête, l'estimation du nombre total de véhicules-kilomètres qui en résulte sera en deçà du total réel pour la population. Ces biais ne sont pas reflétés dans l'estimation de l'erreur-type.

Les erreurs non dues à l'échantillonnage ne sont peut-être qu'une partie de l'erreur d'enquête totale, mais leur contribution peut être importante. Pour limiter l'effet de ce type d'erreur, on applique un programme d'assurance de la qualité à chaque enquête. Ainsi, des suivis en cas de non-réponse peuvent être effectués pour obtenir des renseignements de la part de l'ensemble des non-répondants ou, dans le cas des questionnaires qui ne sont remplis que partiellement, obtenir des réponses aux questions jugées essentielles. Diverses procédures d'assurance de la qualité peuvent être appliquées à l'étape de la saisie des données. Les procédures de vérification des données font ressortir certaines incohérences dans la structure des données et les procédures d'imputation peuvent alors à corriger les incohérences relevées.

En général, les erreurs non dues à l'échantillonnage sont difficiles à quantifier. Il faut procéder à des études spéciales pour les estimer. Cependant, certaines mesures comme les taux de réponse et d'imputation s'obtiennent facilement et peuvent servir d'indicateurs des erreurs non dues à l'échantillonnage. Divers types d'erreurs non dues à l'échantillonnage sont présentés ci-après.

Erreurs de couverture

Il y a erreur de couverture lorsque la population observée ne couvre pas bien la population d'intérêt. Par conséquent, certaines unités appartenant à la population d'intérêt sont soit exclues (sous-dénombrement) soit comptées plus d'une fois (surdénombrement). En outre, des unités qui sont hors du champ d'enquête peuvent être présentes dans la population observée (surdénombrement).

Les sources suivantes d'erreurs de couverture ont été remarquées pour l'EVC :

- Les erreurs dans les variables de classification de l'enquête peuvent donner lieu au sous-dénombrement ou au surdénombrement des véhicules immatriculés.
- L'échantillon est tiré de la liste créée trois mois avant le début de la période de référence. Ainsi, les véhicules immatriculés après la création de la liste et avant la fin de la période de référence ne peuvent pas faire partie de l'échantillon.
- Lorsque la liste des véhicules d'un secteur de compétence n'a pas été créée à temps ou n'est pas arrivée du tout, le sous-dénombrement est encore plus marqué, puisqu'il faut utiliser une vieille liste pour l'échantillonnage.
- Lorsque la liste des véhicules d'un secteur de compétence a été créée à l'avance, il y a surdénombrement.
- Lorsqu'un véhicule a été mis hors service ou destiné à la récupération de pièces et est demeuré sur la liste, il y a surdénombrement.
- La population observée (voir la section « Qualité des données, concepts et méthodologie », Statistique Canada – no 53F0004 au catalogue) peut comprendre des véhicules avec le même Numéro d'Identification de Véhicule (NIV) dans plusieurs provinces. Comme chaque véhicule a un NIV unique, ceci va probablement causer le surdénombrement et par conséquent la surestimation.
- Si un véhicule est immatriculé après la création d'une liste d'immatriculation et que cette immatriculation se termine avant la création de la liste d'immatriculation suivante, le véhicule en question ne fait pas partie d'aucune liste et constitue une source de sous-dénombrement.

Ainsi, l'EVC est sujette à un certain degré de sous-dénombrement et de surdénombrement. La procédure d'estimation sert à compenser la partie du sous-dénombrement et du surdénombrement qui a été déterminée.

Les hypothèses de travail à l'estimation supposent que les réponses fournies par les répondants sont justes à moins que l'on n'ait des preuves sérieuses du contraire. En conséquence, les corrections apportées à l'estimation pour traiter le sous-dénombrement et le surdénombrement s'appuient sur l'ensemble des données fournies par les répondants.

Erreurs de réponse

Il y a erreur de réponse lorsqu'un répondant donne des renseignements erronés parce que les questions de l'enquête ont été mal interprétées ou qu'il n'a pas les bons renseignements, qu'il donne de faux renseignements par erreur, ou qu'il ne veut pas révéler les bons renseignements. La vérification a de bonnes chances de repérer les erreurs de réponse importantes. Cependant, d'autres erreurs de réponse pourraient passer inaperçues.

La vérification des données a fait ressortir peu d'erreurs de réponse.

Erreurs de non-réponse

Il y a erreur de non-réponse lorsqu'un répondant ne répond pas du tout (non-réponse totale) ou ne répond qu'à certaines questions (non-réponse partielle). Ces erreurs peuvent être graves si les caractéristiques des nonrépondants sont systématiquement différentes de celles des répondants et / ou si le taux de non-réponse est élevé. À cet effet, voir les tableaux des taux de réponse dans la section « Qualité des données, concepts et méthodologie », Statistique Canada – no 53F0004 au catalogue).

Erreurs de traitement

Outre les erreurs de couverture, de réponse et de non-réponse qui sont décrites plus haut, les erreurs qui surviennent pendant le traitement des données constituent une autre composante de l'erreur non due à l'échantillonnage. Les erreurs de traitement peuvent survenir pendant la saisie, le codage, la transcription, la vérification, l'imputation, la détection et le traitement des valeurs aberrantes, et d'autres types de manipulation des données.

Il y a erreur de codage lorsqu'un champ est mal codé à cause d'une mauvaise interprétation des procédures de codage ou d'une erreur de jugement (p. ex., erreurs de codage de marchandises). Il y a erreur de saisie lorsque les données sont mal interprétées ou mal entrées. Par exemple, une lecture d'odomètre de 53 467 pourrait être entrée 54 367.

Une fois codées et saisies, les données sont sujettes à la vérification et l'imputation des valeurs manquantes ou erronées. La qualité des données utilisées dans l'estimation dépend de la quantité d'imputation et de la différence entre les valeurs imputées et les valeurs vraies, mais inconnues. S'il repose sur de mauvaises hypothèses ou s'il est incapable de trouver une valeur imputée, le processus d'imputation risque d'introduire un biais dans les estimations. Dans le cas de l'EVC par exemple, il est impossible de détecter une valeur manquante ou erronée de carburant acheté pour les véhicules qui parcourent seulement une petite distance durant la période visée.

Indicateur de qualité

Les taux de réponse par province et type de véhicules varient habituellement entre 50 et 80 %.

Il faut examiner simultanément le c.v., le taux de réponse et le taux d'imputation relatif pour évaluer la fiabilité d'une estimation. Un indicateur de qualité, accompagnant chaque estimation, aide l'utilisateur à évaluer l'effet possible de la non-réponse, de l'imputation et de l'erreur d'échantillonnage. L'indicateur de qualité tient compte simultanément du c.v. et du taux d'imputation relatif.

Indicateur de la qualité	C.V. équivalent	Explication de la qualité de l'estimation
A	Moins de 5 %	Excellente
B	5 % à 10 %	Très bonne
C	10 % à 15 %	Bonne
D	15 % à 20 %	Acceptable
E	20 % à 35 %	À utiliser avec prudence
F	35 % et plus	Trop peu fiable pour être publié

Selon ces mesures, les estimations au niveau national sont considérées généralement de qualité suffisante pour être publiées. Les estimations au niveau provincial sont d'une qualité moindre dans la plupart des cas.