



Guide de l'utilisateur des microdonnées

Enquête internationale auprès des jeunes

2006



Statistique
Canada

Statistics
Canada

Canada

Table des matières

1.0	Introduction	5
2.0	Contexte	7
3.0	Objectifs	9
4.0	Concepts et définitions	11
5.0	Méthodologie de l'enquête	13
5.1	Population visée	13
5.2	Plan de sondage	13
5.2.1	Stratification primaire	13
5.2.2	Sélection de l'échantillon	13
5.2.3	Taille et répartition de l'échantillon	14
6.0	Collecte des données	15
6.1	Conception du questionnaire	15
6.2	Opérations sur le terrain	15
7.0	Traitement des données	19
7.1	Saisie des données	19
7.2	Vérification	19
7.3	Codage des questions ouvertes	19
7.4	Création de variables dérivées	19
7.5	Pondération	21
7.6	Suppression de renseignements confidentiels	21
8.0	Qualité des données	23
8.1	Taux de réponse	23
8.2	Erreurs relatives à l'enquête	23
8.2.1	Base de sondage	24
8.2.2	Collecte des données	24
8.2.3	Traitement des données	25
8.2.4	Non-réponse	25
8.2.5	Mesure de l'erreur d'échantillonnage	26
9.0	Lignes directrices pour la totalisation, l'analyse et la diffusion de données	27
9.1	Lignes directrices pour l'arrondissement d'estimations	27
9.2	Lignes directrices pour la pondération de l'échantillon en vue de la totalisation	28
9.3	Définitions de types d'estimations : catégoriques et quantitatives	28
9.3.1	Estimations catégoriques	28
9.3.2	Estimations quantitatives	28
9.3.3	Totalisation d'estimations catégoriques	29
9.3.4	Totalisation d'estimations quantitatives	29
9.4	Lignes directrices pour l'analyse statistique	30
9.5	Lignes directrices pour la diffusion de coefficients de variation	30
9.6	Seuils pour la diffusion des estimations pour l'Enquête internationale auprès des jeunes	32

10.0	Tables de variabilité d'échantillonnage approximative	33
10.1	Comment utiliser les tables de coefficients de variation pour des estimations catégoriques.....	34
10.1.1	Exemples d'utilisation des tables de coefficients de variation pour des estimations catégoriques	35
10.2	Comment utiliser les tables de coefficients de variation pour obtenir des limites de confiance	39
10.2.1	Exemple d'utilisation des tables de coefficients de variation pour obtenir des limites de confiance	40
10.3	Comment utiliser les tables de coefficients de variation pour effectuer un test t.....	40
10.3.1	Exemple d'utilisation des tables de coefficients de variation pour effectuer un test t	41
10.4	Coefficients de variation pour des estimations quantitatives.....	41
10.5	Tables des coefficients de variation.....	41
11.0	Pondération	43
12.0	Questionnaire	47
13.0	Cliché d'enregistrement à valeurs univariées	49

1.0 Introduction

Statistique Canada a mené l'Enquête internationale auprès des jeunes (EIJ) au printemps 2006 avec le concours et l'appui du Centre national de prévention du crime (CNPC), division du ministère fédéral de la Sécurité publique et Protection civile Canada. Ce manuel a été produit pour faciliter la manipulation du fichier de microdonnées portant sur les résultats de l'enquête.

Toutes les questions concernant l'ensemble de données ou son utilisation devraient être adressées à :

Statistique Canada

Services à la clientèle
Division des enquêtes spéciales
Téléphone : 613-951-3321 ou appelez sans frais : 1-800-461-9050
Télécopieur : 613-951-4527
Courriel : des@statcan.ca

Sécurité publique et Protection civile Canada

Lucie Léonard
Centre national de prévention du crime
22, rue Queen, 12^e étage
Ottawa (Ontario)
K1A 0P8
Téléphone : 613-957-6362
Télécopieur : 613-941-9013
Courriel : Lucie.Leonard@PSEPC-SPPCC.gc.ca

2.0 Contexte

Il y a toujours eu de la part des chercheurs, des décideurs, des éducateurs et du grand public un vif intérêt pour le comportement des jeunes en général et leurs actes d'inconduite en particulier. Les données statistiques puisées à des sources policières sur la délinquance juvénile ne décrivent que les méfaits ou les actes criminels déclarés. Les statistiques policières et judiciaires ne rendent compte que d'une très faible proportion des actes d'inconduite chez les jeunes et nous renseignent très peu sur leurs circonstances familiales ou personnelles. Pour étudier le phénomène de l'inconduite dans le contexte des rapports ou liens avec les parents, les amis et les gens fréquentés à l'école, on se doit d'interroger les jeunes.

Lancée par le Centre de recherche et de documentation du ministère néerlandais de la Justice, l'Étude internationale sur l'autodéclaration de la délinquance (EIAD) a d'abord été réalisée en 1992 dans 13 pays d'Europe et dans l'État du Nebraska aux États-Unis. On y a analysé et interprété des données sur les liens à établir entre la délinquance, l'âge, le sexe et divers facteurs de risque. Pour mieux cerner ces relations, on s'est reporté à certaines mesures de contrôle social en milieu familial et scolaire. On s'est également intéressé au rôle des pairs et des activités de loisir dans le comportement des jeunes. Les résultats de cette première étude ayant été jugés fort intéressants, on a décidé de reprendre l'exercice en 2006 (EIAD 2).

Le Centre national de prévention du crime du ministère fédéral de la Sécurité publique et de la Protection civile Canada a parrainé la participation canadienne à cette réédition de l'étude internationale où on s'est attaché une fois de plus aux comportements et à l'inconduite des élèves de la 7^e à la 9^e année, mais cette fois dans une trentaine de pays, surtout en Europe.

Dans le cas du volet canadien de l'étude désigné par le titre du questionnaire Enquête internationale auprès des jeunes (EIJ), on a donné encore plus de valeur aux résultats en concentrant l'échantillon dans une grande agglomération urbaine pour laquelle on a examiné les données d'enquête avec d'autres renseignements comme les données locales de recensement et les indications sur la criminalité au niveau des quartiers. La ville de Toronto a été retenue comme la ville où Statistique Canada pourrait réaliser l'enquête et sur laquelle portera l'analyse des résultats.

3.0 Objectifs

L'Enquête internationale auprès des jeunes (EIJ) fournit des renseignements détaillés sur les actes d'inconduite des jeunes et répond à d'importantes questions sur les facteurs de risque et de protection qui interviennent, ainsi que sur la façon dont les écoles et les collectivités peuvent aider les enfants à risque à adopter des comportements prosociaux et à obtenir des résultats scolaires positifs.

Cette enquête a réuni des données sur les thèmes suivants :

- Antécédents et liens familiaux;
- Amitiés et activités en période de loisirs;
- Attachement à l'école et au quartier;
- Engagement scolaire mesuré par la réussite et la fréquentation autodéclarées;
- Expériences personnelles et familiales traumatisantes;
- Consommation d'alcool et de drogue;
- Fréquence de divers types de délinquance (vandalisme, vol, violence, utilisation illicite d'Internet, etc.);
- Opinions au sujet du comportement violent chez les jeunes;
- Impulsivité, maîtrise de la colère et penchant pour les comportements risqués (autodéclarés);
- Temps consacré au travail rémunéré et au bénévolat.

Utilisant un questionnaire et une méthode de collecte de données semblables à ceux employés dans d'autres pays, l'EIJ rend possibles les comparaisons de fréquence des types d'inconduite des jeunes dans le monde industrialisé et un examen international de la variabilité de ce qu'on appelle les corrélats des comportements autodéclarés de délinquance.

4.0 Concepts et définitions

Ce chapitre donne un aperçu des concepts et des définitions d'intérêt pour les utilisateurs. Les utilisateurs sont priés de se reporter au chapitre 12.0 de ce document où figure une copie du questionnaire d'enquête réellement employé.

Comportements de délinquance – Il s'agit de tous les comportements qui enfreignent le *Code criminel du Canada* chez les jeunes de 12 à 17 ans. On notera que la consommation d'alcool, de tabac et de drogue est appréhendée par l'enquête, mais considérée non pas comme un comportement de délinquance, mais comme un facteur de risque.

Victimisation – Il s'agit d'actes d'exploitation ou de traitement abusif des personnes. Dans l'Enquête internationale auprès des jeunes (EIJ), on s'intéresse à trois types de victimisation :

- Il y a d'abord les **actes de violence** sous forme de vols qualifiés, de vols simples ou de tentatives de vol dont l'auteur a une arme ou use de violence ou de menaces de violence à l'égard de la victime. Les actes de victimisation avec violence comprennent les voies de fait se manifestant par une agression (la victime est frappée, giflée, empoignée, jetée au sol ou battue), une menace directe de sévices ou la présence d'une arme.
- Il y a aussi les **vols** simples ou tentatives de vol d'effets personnels (argent, cartes de crédit, vêtements, bijoux, sacs à main, portefeuilles, etc.) dont l'auteur, à la différence des auteurs de vols qualifiés, n'affronte pas la victime.
- Il y a enfin les actes d'**intimidation**.

École buissonnière – Il s'agit des absences délibérées de l'école à l'insu et sans le consentement des parents.

Définitions intégrées dans les questions - Le questionnaire EIJ présente quelques définitions pour les questions où l'enquêté a besoin de recevoir des éclaircissements ou d'être dirigé. Voici deux exemples de définition ou d'illustration de concepts dans ce questionnaire :

- Question 15.4 – On t'a intimidé(e) à l'école (d'autres élèves t'ont humilié(e), se sont moqués de toi, t'ont frappé(e) ou donné des coups de pieds, ou t'ont exclu(e) de leur groupe)?
- Question 70 – As-tu déjà fait du piratage informatique (franchi les barrières de sécurité d'un site Web ou d'un compte d'ordinateur)?

Types de délinquance - Dans la principale tranche de questions de l'enquête (questions 48 à 71), on s'enquiert des actes de délinquance des répondants et/ou de leurs amis. On cite trois types de délinquance :

- actes de violence;
- actes contre les biens;
- actes avec Internet (on ajoute un acte de délinquance consistant à vendre de la drogue ou à servir d'intermédiaire).

Les actes de violence sont à la fois les violences physiques et les actes liés aux possibilités de violence. Par cinq questions, on s'interroge sur ce type de délinquance, et, plus précisément, sur les vols de sacs à main et autres vols à la tire, le port d'armes, les menaces armées ou menaces de sévices pour arracher de l'argent, la participation à des bagarres en groupe et enfin le fait de battre ou de blesser.

Les actes contre les biens sont les vols d'effets personnels, les vols dans les magasins, les introductions par effraction, le vandalisme public ou privé et l'incendie volontaire.

Les actes avec Internet sont enfin les téléchargements de piratage, les atteintes à la sécurité des sites Web ou des comptes d'ordinateur, l'envoi de messages de harcèlement et la diffusion de matériel pornographique.

5.0 Méthodologie de l'enquête

L'Enquête internationale auprès des jeunes (EIJ) a été menée à Toronto auprès d'un échantillon de jeunes de la 7^e à la 9^e année. Il s'agissait de jeunes fréquentant les écoles publiques du Conseil scolaire du district de Toronto (CSDT) ou les écoles privées de la région métropolitaine de Toronto en avril et mai 2006.

5.1 Population visée

La population visée était formée d'élèves des 7^e, 8^e et 9^e années fréquentant les écoles publiques du Conseil scolaire du district de Toronto ou les écoles privées dans la région métropolitaine de Toronto au moment où les données ont été recueillies. Il s'agit en gros de 60 000 jeunes. Il convient de noter que le Conseil des écoles catholiques de Toronto a refusé de participer à cette étude et que, par conséquent, ses élèves ne font pas partie de la population cible et ne sont pas représentés dans l'échantillon. On estime que les élèves des écoles catholiques constituent environ le quart de la population scolaire de la région métropolitaine de Toronto. Les jeunes qui ont décroché de l'école ou qui n'y vont pas pour d'autres raisons n'appartiennent pas non plus à la population visée. On a enfin exclu de cette population les jeunes des écoles spéciales.

La population enquêtée diffère très légèrement de la population visée, puisque la clientèle des petites écoles comptant 10 élèves au plus pour toute une année d'études a été exclue. C'est là moins de 0,5 % des élèves dans la population visée. Les intéressés sont proportionnellement plus nombreux dans les écoles privées que dans les écoles publiques (3 % contre moins de 1 %).

5.2 Plan de sondage

5.2.1 Stratification primaire

Il y a trois variables de stratification, à savoir l'année d'études, la région géographique et le type d'école (publique ou privée). Avec trois niveaux de stratification, on obtiendrait des strates trop petites (et des taux élevés d'échantillonnage), aussi a-t-on consulté les utilisateurs pour juger des principaux domaines d'intérêt à des fins d'analyse. La rétroaction recueillie a permis d'opter pour les années d'études et les régions géographiques (au nombre de deux), ce qui a donné six strates. On a délimité les régions d'après les codes postaux de sorte que les populations scolaires soient à peu près les mêmes. L'échantillonnage a été indépendant dans chaque strate et un certain nombre d'écoles ont été choisies plusieurs fois pour des années d'études différentes.

5.2.2 Sélection de l'échantillon

Dans chaque strate, il y a eu sélection systématique des écoles avec probabilité proportionnelle à la taille. La mesure de taille a été l'effectif scolaire pour l'année d'études. C'est l'intervieweur de Statistique Canada qui, sur le terrain, a sélectionné les classes. Le choix s'est fait au hasard pour l'année d'études. L'échantillon final est de 210 classes dans 176 écoles.

5.2.3 Taille et répartition de l'échantillon

L'échantillon était d'une répartition proportionnelle dans les six strates. On a calculé la taille d'échantillon à 3 150 répondants pour un coefficient de variation de 16,5 % ou moins dans chaque domaine d'estimation selon une proportion seuil de 12 %. On a ensuite majoré cette taille en fonction de la non-réponse. En se fondant sur les tailles d'effectif des strates, on a calculé le nombre d'élèves requis dans chacune et, de là, on a estimé le nombre de classes à sélectionner.

6.0 Collecte des données

On a recueilli les données de l'Enquête internationale auprès des jeunes dans les écoles publiques et privées de la ville de Toronto en avril et mai 2006. Les interviews réalisées relevaient des dispositions de participation volontaire de la *Loi sur la statistique*. Il fallait obtenir le consentement exprès des parents ou des tuteurs avant que les élèves ne soient admis à participer. Ceux-ci avaient la possibilité à la fin de ne pas remplir le questionnaire lorsqu'il était distribué par les intervieweurs en début de classe.

6.1 Conception du questionnaire

Le comité directeur de l'étude internationale a produit la version anglaise du questionnaire provisoire (en s'inspirant largement du questionnaire de la première étude de 1992). Les équipes de recherche en milieu scolaire ont traduit et essayé le document dans la plupart des pays participants au printemps et au début de l'été 2005.

Avant le prétest, l'équipe de Statistique Canada a apporté quelques modifications au questionnaire, soucieuse à la fois de préserver l'original et d'en adapter la forme et le fond aux élèves canadiens.

Le prétest a été réalisé par une maison d'études de marché en août 2005. Les 34 participants choisis représentaient les années d'études (de la 7^e à la 9^e), les sexes, les langues d'enseignement (français et anglais), l'intérieur et l'extérieur de la région urbaine d'Ottawa-Gatineau et les résultats scolaires (autodéclarés) moyens et supérieurs et inférieurs à la moyenne.

On a réparti les participants en six groupes (garçons et filles ont été interviewés séparément). Chaque séance en groupe a duré environ deux heures. Les participants ont d'abord reçu des explications sur la façon dont l'enquête serait menée dans les écoles. Ils ont ensuite rempli le questionnaire sur papier. Ils ont enfin fait part de leurs commentaires et répondu aux questions de rétroaction de l'animateur.

À la suite des commentaires reçus de l'équipe de Statistique Canada et des résultats du prétest, on a proposé des modifications au questionnaire international à l'occasion d'une réunion des pays sur cet instrument d'enquête en 2005.

La version anglaise définitive distribuée par le comité directeur international tient compte de la plupart des recommandations de Statistique Canada, mais non de la totalité.

Pour cette version, il a fallu apporter des modifications de langue et de grammaire et de légers changements de formulation et de présentation des questions. On a essayé le questionnaire une fois de plus, officieusement cette fois, auprès d'un petit groupe d'élèves de la région d'Ottawa-Gatineau. Il y a eu d'autres modifications, mais il a fallu les limiter pour ne pas nuire à la comparabilité des résultats.

6.2 Opérations sur le terrain

L'enquête dans les écoles a eu lieu de février à mai 2006. On a notamment posté une lettre de présentation aux écoles publiques et privées de l'échantillon de Toronto, sélectionné les classes pour participer à l'enquête et mené les séances où les élèves ont répondu au questionnaire sur papier. Les activités de collecte ont été précédées d'une longue démarche engagée en septembre 2005 pour l'obtention de l'approbation des autorités scolaires.

Voici en bref ce qu'ont été les activités de collecte :

Prise de contact avec l'école

Peu de temps après que le bureau régional eût posté la lettre de présentation aux écoles sélectionnées en février 2006, les intervieweurs ont téléphoné aux diverses écoles pour prendre contact et solliciter la collaboration du directeur ou de l'administrateur scolaire.

Si celui-ci refusait de participer, l'école n'était pas remplacée, mais on mettait tout en œuvre pour amener le directeur ou l'administrateur à consentir à l'enquête.

Première visite à l'école

À son arrivée à l'école, l'intervieweur se présentait au directeur et décrivait sommairement les activités de collecte. Il se servait d'une formule étiquetée de sélection de classe à des fins de contrôle. Sur cette formule, il indiquait l'année d'études choisie. Si l'école avait plusieurs classes pour cette année, il se reportait à la grille de sélection sur l'étiquette pour choisir une de ces classes au hasard.

Il énumérait les élèves de la classe choisie en précisant le nom et le sexe, si celui-ci était connu. Pour chaque intéressé, il préparait une trousse à emporter à la maison avec une lettre de présentation, une formule de consentement des parents et une lettre s'adressant directement à l'élève sélectionné. Comme les nouveaux immigrants sont nombreux à Toronto, il prévoyait une fiche d'information multilingue pour les parents qui pouvaient avoir de la difficulté à communiquer en français ou en anglais.

Le directeur ou l'enseignant était prié de remettre les formules de consentement des parents et d'en contrôler la réception une fois remplies. L'intervieweur indiquait qu'il repasserait au moment fixé pour reprendre les formules de consentement remplies.

Deuxième visite à l'école

L'intervieweur revenait à l'école pour reprendre les formules et organiser la séance en classe.

Il demandait à l'enseignant ou à la personne-ressource de distribuer une lettre de rappel aux parents n'ayant pas encore renvoyé la formule de consentement. Il joignait un exemplaire de la formule à cette lettre.

Le Conseil scolaire du district de Toronto n'a malheureusement pas autorisé la collecte des numéros de téléphone des parents et, par conséquent, on a dû passer par les administrateurs scolaires pour faire le suivi des formules de consentement. Si les formules n'étaient pas signées par les parents ou remises par les élèves, ceux-ci ne pouvaient participer à l'enquête.

Séance en classe (troisième visite à l'école)

L'intervieweur a préparé un questionnaire pour chaque enfant admissible. À la page couverture, il reportait le numéro d'identification de l'élève provenant de la formule de sélection de classe. Ce numéro était formé du numéro de l'école, de l'année d'études et d'un numéro attribué arbitrairement à l'élève. L'intervieweur n'écrivait pas le nom de l'élève sur le questionnaire par souci de respecter l'anonymat. Le nom des élèves ne figurait que sur les enveloppes contenant les questionnaires à leur remettre. On reprenait les questionnaires remplis sans les enveloppes.

Une fois en classe, l'intervieweur suivait la procédure suivante :

- il se présentait aux élèves;
- il exposait l'objet de l'enquête;
- il demandait à l'enseignant de distribuer les enveloppes avec des portemines gratuits;
- il lisait la note de présentation à haute voix au début du questionnaire;
- il répondait à quelques questions déterminées d'avance avec les élèves pour leur montrer comment remplir différentes parties;
- il disait aux élèves de ne pas hésiter à lever la main pour poser discrètement leurs questions;
- il reprenait d'abord les enveloppes portant le nom des élèves et ensuite les questionnaires remplis; il mettait ceux-ci dans un Versapak (sac souple sécurisé) pour les transporter;
- il remerciait les élèves et l'enseignant de leur collaboration et leur appui.

La séance en classe durait en moyenne de 40 à 50 minutes. On demandait à l'enseignant de rester en classe, mais de ne pas circuler parmi les élèves afin de préserver l'anonymat et la confidentialité des réponses.

Les intervieweurs devaient porter les indications suivantes sur la formule de sélection de classe :

- nombre de classes pour l'année d'études visée;
- nombre d'élèves de l'ensemble des classes pour l'année d'études visée;
- nombre d'élèves de la classe choisie et nombre de garçons et de filles;
- nombre d'élèves ayant ou n'ayant pas renvoyé la formule de consentement;
- nombre de formules renvoyées avec ou sans le consentement écrit des parents.

On s'est ensuite reporté à une partie de cette information pour calculer les taux de réponse et de non-réponse.

Il y a également une partie de la formule de sélection de classe où on a donné en particulier des indications directement demandées par l'organe de coordination internationale :

- type de personnel scolaire présent (maître ou titulaire de classe, enseignant spécialisé ou autre);
- sexe du personnel présent;
- nombre d'observateurs présents (de 0 à 2 au total);
- nombre d'élèves absents ce jour-là;
- durée de la séance en classe en minutes de bout en bout.

7.0 Traitement des données

Le principal produit de l'Enquête internationale auprès des jeunes (EIJ) est un fichier de microdonnées « épuré ». Ce chapitre présente un bref résumé des phases de traitement inhérentes à la production de ce fichier.

7.1 Saisie des données

Il y a eu saisie des données de tous les questionnaires de l'EIJ par imagerie numérique au bureau central de Statistique Canada à Ottawa. De plus, on a entré manuellement les réponses « autre, préciser » qui ne pouvaient être reconnues par imagerie numérique. On a vérifié intégralement les numéros d'identification des élèves pour prévenir toute erreur d'entrée manuelle. On a contrôlé la qualité de la saisie des données par une opération de vérification au hasard portant sur presque 20 % de cette tâche. Le taux d'erreur était de moins de 1 %.

7.2 Vérification

On a conçu le questionnaire de l'Enquête internationale auprès des jeunes en prévoyant très peu de sauts ou d'enchaînements, jugeant que ceux-ci seraient mal respectés par des répondants plus jeunes.

L'équipe d'enquête a pris la décision de vérifier les questionnaires à la fois de haut en bas et de bas en haut. Pour accomplir cette tâche, il fallait déterminer les enchaînements avant même d'établir les programmes de vérification.

Les erreurs dans le déroulement du questionnaire, où l'on a relevé des questions qui ne s'appliquaient pas au répondant (et auxquelles on n'aurait donc pas dû répondre) et qui renfermaient des réponses, constituaient le premier type d'erreurs traitées. Dans ces cas, une vérification par ordinateur a éliminé automatiquement les données superflues en suivant l'ordre du questionnaire dicté par les réponses à des questions antérieures et subséquentes, parfois.

Le second type d'erreurs traitées avait trait à un manque d'information dans les questions pour lesquelles le répondant aurait dû répondre. Pour ce type d'erreur, un code de non-réponse ou « non déclaré » était attribué au poste.

Voici les codes qu'utilise normalement Statistique Canada à cette fin :

6, 96, 996, etc. = Saut ou enchaînement valide

7, 97, 997, etc. = Ne sait pas

8, 98, 998, etc. = Refus

9, 99, 999, etc. = Non déclaré

7.3 Codage des questions ouvertes

Le questionnaire de l'EIJ comprenait huit questions partiellement ouvertes avec une catégorie de réponse « autre » ou « autre - préciser » pouvant contenir une réponse écrite. À la question 3.1 par exemple, il devait préciser dans quel pays il était né s'il ne s'agissait pas du Canada. Dans ce cas, on s'est reporté au codage normalisé de pays de naissance pour coder les réponses lisibles que donnaient les répondants à la question.

7.4 Création de variables dérivées

Un certain nombre de données élémentaires incluses dans le fichier de microdonnées ont été calculées en combinant des postes sur le questionnaire pour faciliter l'analyse des données.

Dans chaque cas, on indique dans le livre des codes avec quelles questions on a créé la variable et, si on a calculé des valeurs de résultat, comment le calcul s'est fait. On trouvera les variables dérivées au cliché d'enregistrement à la suite des questions EIJ avec lesquelles elles ont été créées.

Il en existe deux types au fichier. Les variables créées avec plusieurs questions ou postes du questionnaire sont ordinaires et leur désignation commence par « D » pour variable dérivée. Par ailleurs, si on regroupe ou fond les réponses à une question en moins de catégories, la variable en question sera désignée par la lettre « G » pour « variable groupée ».

Voici des exemples de variables dérivées et groupées :

Désignation de la variable : DYOTHSES

Questions utilisées : Q11 As-tu ta propre chambre à coucher?
Q12 As-tu à la maison un ordinateur que tu as la permission d'utiliser?
Q13 As-tu ton propre téléphone cellulaire?
Q14 Ta famille a-t-elle une voiture?

Description : Cette variable dérivée réunit quatre questions par « Oui/Non » en une variable unique permettant d'évaluer approximativement la situation socio-économique (SSE) du ménage de l'élève. Voici les catégories de réponse créées pour cette variable :

- 1 Possède/accès à aucun(e), 1 ou 2 items
- 2 Possède/accès à 3 ou 4 items
- 9 Non déclaré

Désignation de la variable : DVICTSCR

Questions utilisées : Q15 Réfléchis aux 12 derniers mois et indique si tu as déjà eu l'un ou l'autre des problèmes ci-après?

Q15_1A	Quelqu'un a voulu te forcer à lui donner de l'argent...
Q15_2A	Quelqu'un t'a frappé(e) violemment...
Q15_3A	Quelqu'un t'a volé quelque chose...
Q15_4A	On t'a intimidé(e) à l'école...

Description : Cette variable dérivée de la victimisation indique le nombre de types de victimisation qu'a subis le répondant au cours des 12 derniers mois. C'est la somme des réponses par l'affirmative aux quatre questions énumérées plus haut. Les valeurs vont de 0 à 4.

Désignation de la variable : G17

Question utilisée : Q17 De quelle façon t'entends-tu en général avec la femme avec qui tu vis (ta mère ou ta belle-mère)?

Description : Cette variable groupée fait seulement intervenir les réponses à Q17. Les catégories et valeurs initiales de réponse

- 1 Je m'entends très bien
- 2 Je m'entends assez bien
- 3 Je ne m'entends pas très bien
- 4 Je ne m'entends pas du tout

On a conservé les catégories de réponse 1 et 2, tandis que les catégories 3 et 4 ont été groupées afin de former la nouvelle catégorie de réponse groupée.

- 3 Je ne m'entends pas très bien/pas du tout

7.5 Pondération

Le principe qui sous-tend une estimation pour un échantillon probabiliste veut que chacune des personnes incluses dans l'échantillon « représente », en plus d'elle-même, plusieurs autres personnes qui en sont exclues. Par exemple, dans un échantillon aléatoire simple de 2 % de la population, chaque personne incluse dans l'échantillon représente 50 membres de la population.

La phase de la pondération est une étape où l'on calcule ce nombre (ou poids) pour chaque enregistrement. Ce poids, qui figure dans le fichier de microdonnées, **doit** servir à calculer des estimations significatives à partir de l'enquête. Si, par exemple, le nombre d'élèves de la 8^e année qui ont déjà bu de la bière, des panachés (« coolers ») ou du vin (Q49), doit être estimé, cette opération s'effectue en sélectionnant les enregistrements se référant aux élèves inclus à l'intérieur de l'échantillon qui présentent cette caractéristique et en additionnant les poids inscrits dans ces enregistrements.

Le chapitre 11.0 renferme des détails au sujet de la méthode utilisée pour calculer ces poids.

7.6 Suppression de renseignements confidentiels

Il convient de souligner que les fichiers de microdonnées « à grande diffusion » (FMGD) peuvent différer des fichiers « maîtres » de l'enquête que conserve Statistique Canada. Ces différences sont habituellement le résultat de mesures prises pour protéger l'anonymat des répondants à une enquête. Les mesures les plus courantes sont la suppression de variables du fichier, le regroupement de valeurs en des catégories plus étendues et le codage de valeurs spécifiques à la catégorie « non déclaré ». Les utilisateurs ayant besoin d'avoir accès à de l'information exclue des fichiers de microdonnées peuvent acheter des totalisations spéciales. Les estimations produites seront communiquées à l'utilisateur, sous réserve du respect des lignes directrices pour l'analyse et la diffusion dont le chapitre 9.0 de ce document fournit un aperçu.

Le fichier maître de données de l'enquête inclut plusieurs variables qui ont été retranchées du FMGD de l'EIJ, parce que donnant la possibilité d'identifier des répondants. Il s'agit notamment de l'âge du répondant, de la situation d'immigrant, du pays de naissance, de la composition de la famille, de la langue parlée à la maison, de la situation d'emploi des parents et du redoublement d'années d'études.

Comme exemple de regroupement de valeurs, il y a l'âge des étudiants, lequel est demandé au début de chaque question sur les actes d'inconduite (questions 49 à 71). L'âge des étudiants a été groupé en deux catégories : « Moins de 10 ans » et « 10 ans et plus ».

Pour certaines variables sensibles au risque d'identification d'individus, le FMGD peut avoir subi de la suppression locale, c'est-à-dire que certaines valeurs présentes sur le fichier maître peuvent avoir été codées comme « non déclarées » sur le FMGD. On compte 55 de ces suppressions qui touchent 12 variables.

8.0 Qualité des données

8.1 Taux de réponse

Le tableau qui suit renferme un résumé des taux de réponse au questionnaire de l'Enquête internationale auprès des jeunes (EIJ).

Année d'études	Classes sélectionnées (admissibles)	Classes répondantes	Taux de réponse des classes* (%)	Nombre d'élèves des classes répondantes	Élèves répondants	Taux de réponse des élèves** (%)	Taux global de réponse*** (%)
7 ^e année	70	62	88,6	1 640	1 207	73,6	65,2
8 ^e année	69	58	84,1	1 519	1 148	75,6	63,5
9 ^e année	66	57	86,4	1 338	935	69,9	60,4
Total	205	177	86,3	4 497	3 290	73,2	63,2

Note : Sur 210 classes sélectionnées, 5 ont été jugées inadmissibles, l'école n'ayant pas l'année d'études pour laquelle la classe était choisie.

- * Le taux de réponse des classes est le nombre de classes répondantes exprimé sous forme de pourcentage du nombre de classes sélectionnées jugées admissibles.
- ** Le taux de réponse des élèves est le nombre d'élèves répondants exprimé sous forme de pourcentage du nombre d'élèves des classes répondantes.
- *** Le taux global de réponse est le taux de réponse des classes multiplié par le taux de réponse des élèves.

8.2 Erreurs relatives à l'enquête

Les estimations calculées à partir de cette enquête reposent sur un échantillon d'élèves. Des estimations légèrement différentes auraient pu être obtenues si un recensement complet avait été effectué en reprenant le même questionnaire et en faisant appel aux mêmes intervieweurs, superviseurs, méthodes de traitement, etc. que ceux effectivement utilisés dans l'enquête. L'écart entre les estimations découlant de l'échantillon et celles que donnerait un dénombrement complet réalisé dans des conditions semblables est appelé erreur d'échantillonnage de l'estimation.

Des erreurs qui ne sont pas liées à l'échantillonnage peuvent se produire à presque toutes les étapes des opérations d'enquête. Les intervieweurs peuvent avoir mal compris les instructions, les répondants peuvent se tromper en répondant aux questions, les réponses peuvent être mal saisies sur le questionnaire et des erreurs peuvent survenir lors du traitement et de la totalisation des données. Ces erreurs sont toutes des exemples d'erreurs non dues à l'échantillonnage.

Sur un grand nombre d'observations, les erreurs aléatoires auront peu d'effet sur les estimations calculées à partir de l'enquête. Toutefois, les erreurs systématiques contribuent à biaiser les estimations de l'enquête. Énormément de temps et d'efforts ont été consacrés à réduire les erreurs non dues à l'échantillonnage dans l'enquête. Des mesures d'assurance de la qualité ont été prises à chacune des étapes du cycle de collecte et de traitement des données afin de contrôler la qualité des données. Ces mesures comprenaient le recours à des intervieweurs

hautement qualifiés, une formation poussée des intervieweurs concernant les procédures d'enquête et du questionnaire, l'observation des intervieweurs en vue de cerner les problèmes liés à la conception du questionnaire ou à une mauvaise compréhension des instructions, des procédures visant à s'assurer que les erreurs de saisie des données étaient réduites au minimum ainsi que des vérifications de la qualité du codage et de contrôle ayant pour but d'attester la logique du traitement.

8.2.1 Base de sondage

Dans le cas des écoles publiques, le Conseil scolaire du district de Toronto (CSDT) a fourni à Statistique Canada deux fichiers administratifs qui ont servi à créer la base de sondage. Le premier fichier décrivait les inscriptions par année d'études intermédiaires, soit les 7^e et 8^e années. Compte tenu de la provenance et du caractère à jour des données, on a considéré que cette base était d'une meilleure qualité que toute autre source d'information disponible. Le second fichier décrivait les inscriptions aux écoles secondaires par âge. Comme il n'y avait pas de données de dénombrement par année d'études au fichier des écoles secondaires, l'âge a servi de variable de substitution pour la 9^e année d'études. Dans le cas des écoles privées, on n'avait pas de conseil scolaire auquel demander une bonne information à jour. Statistique Canada a créé une base en consultant les sources d'information publique à jour et les bases plus anciennes dont il disposait.

Le CSDT a fourni le fichier des écoles publiques à l'automne 2006 pour l'année scolaire en cours. Ces données étaient donc les plus actuelles possible. On a jugé que l'information relative aux 7^e et 8^e années d'études était de haute qualité. Dans le cas de la 9^e année, les données n'étaient pas d'aussi haute qualité, mais demeuraient d'une très bonne qualité. Sur les 185 classes sélectionnées dans les écoles publiques, 5 ont été considérées sur le terrain comme inadmissibles (l'école n'avait pas l'année d'études pour laquelle la classe était choisie). Par ailleurs, aucune des 25 classes sélectionnées dans les écoles privées n'était inadmissible.

8.2.2 Collecte des données

Seuls des intervieweurs aguerris de Statistique Canada ont travaillé à cette enquête. La formation des intervieweurs consistait à lire le Manuel de l'intervieweur et comme on voulait être sûr qu'ils connaîtraient bien les concepts et les procédures de l'EIJ, on a réuni en mars 2006 tous les intervieweurs, principaux et subalternes, pour une journée de formation au bureau régional de Toronto. À cette occasion, des représentants du bureau central ont notamment fait des présentations et organisé des exercices.

À l'étape de la collecte des données, les intervieweurs principaux étaient chargés de soutenir et de surveiller les intervieweurs subalternes. Les membres de l'équipe de projet se sont rendus à Toronto pour observer plusieurs séances en classe. Dans presque toutes les séances qu'ils ont observées, tant les intervieweurs que les enseignants ont agi comme on s'y attendait.

En début de séance, les intervieweurs exposaient l'objet de l'enquête et son caractère confidentiel. Ils disaient clairement que les réponses des élèves seraient protégées et ne seraient donc communiquées à quiconque à l'école ni aux parents. Les questionnaires remplis étaient mis dans un Versapack fourni par Statistique Canada et emportés hors de l'école par l'intervieweur.

La plupart des élèves se sont bien comportés et ont rempli le questionnaire en toute diligence. Ils voulaient que leurs réponses soient précises et ils ont demandé des éclaircissements s'ils avaient du mal à interpréter une question ou des problèmes avec

leur anglais (le tiers des intéressés étaient originaires de l'étranger). À certaines séances, les élèves étaient assis trop près les uns des autres et certains n'ont pas respecté la confidentialité des réponses de leurs voisins.

Les séances en classe ont duré de la fin de mars au milieu de mai 2006.

8.2.3 Traitement des données

On a saisi les données des questionnaires de l'EIJ par imagerie numérique. On a vérifié intégralement les numéros d'identification des élèves. On a aussi choisi au hasard 20 % des réponses à des fins de vérification. Le taux d'erreur était de moins de 1 %. Pour être conservé au fichier, un enregistrement non vierge devait comporter des réponses à quatre questions précises sur cinq (Q6, Q16, Q44, Q50 et Q55). Il n'y a que 55 enregistrements qu'on ait dû écarter.

Comme les jeunes élèves devaient eux-mêmes remplir le questionnaire, on n'y a pas directement intégré de sauts ou d'enchaînements. Pour certaines questions, il y avait une indication graphique, une flèche vers une question à laquelle répondre à la suite d'une première réponse. Dans le cadre de la vérification, on a procédé de bas en haut là où l'élève ne répondait pas à la question initiale, mais répondait à la question de suivi.

Soucieux de préserver les réponses initiales des élèves pour les comparaisons internationales, on n'a pas effectué de contrôles de cohérence ni éliminé les valeurs aberrantes.

8.2.4 Non-réponse

L'une des principales sources d'erreurs non dues à l'échantillonnage observées dans le cadre des enquêtes est l'effet de la non-réponse sur les résultats de l'enquête. L'étendue d'une non-réponse varie d'une non-réponse partielle (le fait de ne pas répondre simplement à une ou à des questions) à une non-réponse complète.

Dans toute l'enquête, il y a eu deux niveaux généraux de non-réponse. D'abord, on a pu constater un certain degré de non-réponse au niveau des écoles (ce qu'expriment les classes répondantes au tableau à la section 8.1). La raison principale en est que le directeur d'école n'a pas consenti à l'enquête. Dans quelques cas, on avait sélectionné une classe, mais sans que les élèves participent ensuite à l'enquête. Il y a aussi eu non-réponse au niveau des élèves. On s'est alors reporté au nombre d'élèves admissibles sur la formule de sélection de classe pour chacune des classes participantes. Cette non-réponse peut s'expliquer par plusieurs facteurs : l'absence de consentement des parents, le refus de participation des élèves, l'absence de l'élève le jour de la collecte, le questionnaire rempli ne contenant pas une information suffisante et ne pouvant être considéré comme utilisable.

On a établi le taux de non-réponse complète par repondération des répondants en compensation des non-répondants.

Dans la plupart des cas, il y a eu non-réponse partielle au questionnaire d'enquête lorsque le répondant n'a pas compris ou a mal interprété une question, a refusé d'y répondre ou ne pouvait se rappeler l'information demandée. Avec un questionnaire à remplir soi-même avec papier et crayon, une partie de la non-réponse partielle est causée par l'inattention ou le manque d'intérêt.

Le taux de non-réponse a été élevé pour les trois questions suivantes (c.-à-d. réponses « Ne sait pas » ou « Non déclaré ») :

- Q36 Les gens diffèrent souvent entre eux quant à leur origine, leur religion et leurs croyances. Tes parents approuvent-ils le fait d'avoir des amis d'un groupe (ethnique) différent? 14 %;
- Q46 Quel niveau de scolarité penses-tu atteindre? 10 %;
- Q71.1 As-tu déjà téléchargé de la musique ou des films d'Internet? Pensaistu que c'était illégal (piratage)? 27 %

8.2.5 Mesure de l'erreur d'échantillonnage

Puisqu'il est inévitable que des estimations établies à partir d'une enquête-échantillon (ou par sondage) soient sujettes à une erreur d'échantillonnage, une saine pratique de la statistique exige que les chercheurs fournissent aux utilisateurs une certaine indication de l'importance de cette erreur d'échantillonnage. Cette section de la documentation renferme un aperçu des mesures de l'erreur d'échantillonnage dont Statistique Canada se sert couramment et dont le Bureau conseille vivement aux utilisateurs qui produisent des estimations à partir de ce fichier de microdonnées à employer également.

La base pour mesurer l'importance potentielle des erreurs d'échantillonnage est l'erreur-type des estimations calculées à partir des résultats d'une enquête.

En raison, cependant, de la diversité des estimations pouvant être produites à partir d'une enquête, l'erreur-type d'une estimation est habituellement exprimée en fonction de l'estimation à laquelle elle se rapporte. La mesure résultante, appelée coefficient de variation (CV) d'une estimation, s'obtient en divisant l'erreur-type de l'estimation par l'estimation elle-même et s'exprime en pourcentage de l'estimation.

Par exemple, supposons que, d'après les résultats de l'enquête, l'on estime que 8,0 % des élèves de la 7^e à la 9^e année ont fumé une cigarette entière et on constate que l'erreur-type de cette estimation est de 0,006.

$$\left(\frac{0,006}{0,080} \right) \times 100 \% = 7,5 \%$$

De plus amples renseignements sur le calcul du coefficient de variation, se trouvent au chapitre 10.0.

9.0 Lignes directrices pour la totalisation, l'analyse et la diffusion de données

Ce chapitre de la documentation renferme un aperçu des lignes directrices que doivent respecter les utilisateurs qui totalisent, analysent, publient ou autrement diffusent des données calculées à partir des fichiers de microdonnées de l'enquête. Ces lignes directrices devraient permettre aux utilisateurs de microdonnées de produire les mêmes chiffres que ceux produits par Statistique Canada, tout en étant en mesure d'obtenir des chiffres actuellement inédits de façon conforme à ces lignes directrices établies.

9.1 Lignes directrices pour l'arrondissement d'estimations

Afin que les estimations qui sont destinées à la publication ou à toute autre forme de diffusion qui sont calculées à partir de ces fichiers de microdonnées correspondent à celles produites par Statistique Canada, nous conseillons vivement 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 normale. 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 normale 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 totaux partiels marginaux et les totaux marginaux des tableaux statistiques doivent être calculés à partir de leurs composantes non arrondies correspondantes, puis ensuite être arrondis à leur tour à la centaine près à l'aide de la technique d'arrondissement normale.
- c) Les moyennes, les proportions, les taux et les pourcentages doivent être calculés à partir de composantes non arrondies (c'est-à-dire des numérateurs et/ou des dénominateurs), puis être arrondis à leur tour à une décimale à l'aide de la technique d'arrondissement normale. Dans le cas d'un arrondissement normal 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.
- d) Les sommes et les différences d'agrégats (ou de rapports) 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) à l'aide de la technique d'arrondissement normale.
- e) Dans les cas, où, en raison de limitations d'ordre technique ou de toutes autres limites, une technique d'arrondissement autre que la technique normale est utilisée produisant des estimations à être publiées ou autrement diffusées différentes des estimations correspondantes publiées par Statistique Canada, nous conseillons vivement aux utilisateurs d'indiquer la raison de ces différences dans le ou les documents à publier ou à diffuser.
- 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.

9.2 Lignes directrices pour la pondération de l'échantillon en vue de la totalisation

Le plan d'échantillonnage utilisé pour l'Enquête internationale auprès des jeunes (EIJ) n'était pas autopondéré. Lorsqu'ils produisent des estimations simples, y compris des tableaux statistiques ordinaires, les utilisateurs doivent appliquer le poids d'enquête approprié.

Si les poids appropriés ne sont pas utilisés, les estimations calculées à partir des fichiers de microdonnées ne peuvent être considérées comme représentatives de la population visée par l'enquête et ne correspondront pas à celles produites par Statistique Canada.

Les utilisateurs devraient également prendre note que certains progiciels pourraient peut-être ne pas permettre la production d'estimations correspondant exactement à celles qu'offre Statistique Canada, en raison du mode de traitement du poids par ces progiciels.

9.3 Définitions de types d'estimations : catégoriques et quantitatives

Avant de discuter de la façon dont on peut totaliser et analyser les données de l'Enquête internationale auprès des jeunes, il est utile de décrire les deux principaux types d'estimations ponctuelles des caractéristiques de la population qui peuvent être produites à partir du fichier de microdonnées créé pour l'EIJ.

9.3.1 Estimations catégoriques

Les estimations catégoriques sont des estimations du nombre ou du pourcentage de membres de la population visée par l'enquête possédant certaines caractéristiques ou faisant partie d'une catégorie définie. Le nombre d'élèves qui ont déjà bu de la bière, des panachés (coolers) ou du vin ou la proportion d'élèves qui aimaient beaucoup l'école en général constitue des exemples de telles estimations. Une estimation du nombre de personnes possédant une certaine caractéristique peut aussi être désignée comme une estimation d'un agrégat.

Exemples de questions catégoriques :

Q : As-tu déjà bu de la bière, des panachés (coolers) ou du vin?
R : Oui / Non

Q : En général, aimes-tu l'école?
R : J'aime beaucoup l'école / J'aime passablement l'école / Je n'aime pas beaucoup l'école / Je n'aime pas du tout l'école

9.3.2 Estimations quantitatives

Les estimations quantitatives sont des estimations de totaux ou de moyennes, de médianes et d'autres mesures d'une tendance centrale de quantités reposant sur certains ou sur tous les membres de la population visée par l'enquête. Elles comprennent aussi expressément des estimations de la forme \hat{X} / \hat{Y} où \hat{X} est une estimation de la quantité totale pour la population visée par l'enquête et \hat{Y} , est une estimation du nombre de personnes dans la population visée par l'enquête qui contribuent à cette quantité totale.

Un exemple d'estimation quantitative est le nombre moyen d'amis qui ont déjà volé quelque chose dans un magasin. Le numérateur est une estimation du nombre total d'amis qui ont déjà volé quelque chose dans un magasin et le dénominateur est le nombre d'élèves ayant déclaré avoir des amis qui ont déjà volé quelque chose dans un magasin.

Exemples de questions quantitatives :

Q: As-tu des amis qui ont déjà volé quelque chose dans un magasin?

R: |_|_| amis

Q: As-tu déjà bu de la bière, des panachés (coolers) ou du vin? As-tu consommé ces boissons au cours des quatre dernières semaines?

R: |_|_| fois

9.3.3 Totalisation d'estimations catégoriques

On peut obtenir des estimations du nombre de gens possédant une certaine caractéristique à partir du fichier de microdonnées en additionnant les poids finals de tous les enregistrements possédant la ou les caractéristiques qui nous intéressent. On obtient des proportions et des rapports de la forme \hat{X} / \hat{Y} en :

- additionnant les poids finals des enregistrements présentant la caractéristique qui nous intéresse pour le numérateur (\hat{X}),
- additionnant les poids finals des enregistrements présentant la caractéristique qui nous intéresse pour le dénominateur (\hat{Y}), puis en
- divisant l'estimation a) par celle de b) (\hat{X} / \hat{Y}).

9.3.4 Totalisation d'estimations quantitatives

On peut obtenir des estimations de quantités à partir du fichier de microdonnées en multipliant la valeur de la variable qui nous intéresse par le poids final de chaque enregistrement, puis en additionnant cette quantité pour tous les enregistrements qui nous intéressent. Pour obtenir, par exemple, une estimation du nombre total de fois que les élèves ont bu de la bière, des panachés (coolers) ou du vin au cours des quatre dernières semaines, multipliez la valeur déclarée à la question Q49_3A (le nombre de fois consommé de la bière, des panachés (coolers) ou du vin au cours des quatre dernières semaines) par le poids final de l'enregistrement, puis additionnez cette valeur pour tous les enregistrements où la variable Q49_3A < 21.

Pour obtenir une moyenne pondérée de la forme \hat{X} / \hat{Y} , le numérateur (\hat{X}) est calculé comme une estimation quantitative et le dénominateur (\hat{Y}) est calculé comme une estimation catégorique. Pour estimer, par exemple, le nombre moyen de fois que les élèves ont bu de la bière, des panachés (coolers) ou du vin au cours des quatre dernières semaines,

- estimez le nombre total de fois que les élèves ont bu de la bière, des panachés (coolers) ou du vin au cours des quatre dernières semaines (\hat{X}) tel qu'il est décrit ci-dessus,
- estimez le nombre d'élèves (\hat{Y}) inclus dans cette catégorie en additionnant les poids finals de tous les enregistrements où la variable Q49_3A < 21, puis

c) divisez l'estimation a) par l'estimation b) (\hat{X} / \hat{Y}).

9.4 Lignes directrices pour l'analyse statistique

L'EIJ repose sur un plan d'échantillonnage complexe comportant une stratification, de multiples étapes de sélection ainsi que des probabilités inégales de sélection des répondants. L'utilisation des données provenant d'enquêtes aussi complexes présente des problèmes pour les analystes, parce que le plan d'enquête et les probabilités de sélection influent sur les procédures d'estimation et de calcul de la variance qui devraient être utilisées. Il faut utiliser les poids de l'enquête pour que les estimations et les analyses des données de l'enquête soient exemptes de biais.

Bien que de nombreuses procédures d'analyse que l'on trouve à l'intérieur de progiciels statistiques permettent d'utiliser des poids, la signification ou la définition du poids inclus dans ces procédures peut différer de ce qui convient dans le contexte d'une enquête-échantillon, de telle sorte que dans bien des cas les estimations produites au moyen de ces progiciels sont correctes, mais que les variances calculées sont piètres. Les variances approximatives pour des estimations simples comme des totaux, des proportions et des rapports (pour des variables qualitatives) peuvent être calculées à partir des tables de variabilité d'échantillonnage approximative qui accompagnent les données.

Pour d'autres techniques d'analyse (de régression linéaire, de régression logistique et de l'analyse de variance, par exemple), il existe une méthode qui peut rendre les variances calculées par l'application des progiciels normalisés plus significatives, en intégrant les probabilités inégales de sélection. L'application de cette méthode entraîne une remise à l'échelle des poids de façon à ce que le poids moyen soit de 1.

Supposons, par exemple, qu'il faut effectuer l'analyse de tous les élèves de sexe masculin. Les étapes à suivre pour remettre à l'échelle les poids sont les suivantes :

- 1) sélectionner tous les élèves du fichier qui ont déclaré Q01 = homme;
- 2) calculer le poids MOYEN pour ces enregistrements en additionnant les poids originaux des élèves établis à partir du fichier de microdonnées pour ces enregistrements puis diviser cette somme par le nombre d'élèves ayant déclaré Q01 = homme;
- 3) pour chacun de ces élèves, calculer un poids REMIS À L'ÉCHELLE égal au poids original de l'élève divisé par le poids MOYEN;
- 4) effectuer l'analyse portant sur ces élèves en utilisant le poids REMIS À L'ÉCHELLE.

Parce qu'on ne tient toujours compte ni de la stratification ni des grappes du plan d'échantillonnage, les estimations des variances calculées avec cette méthode risquent cependant d'être des sous-estimations.

Il faut connaître les détails du plan d'enquête pour calculer des estimations des variances plus précises. De tels détails ne peuvent être fournis dans le fichier de microdonnées en raison de la confidentialité. Statistique Canada peut, contre remboursement des frais, calculer des variances qui tiennent compte du plan complet d'échantillonnage pour beaucoup de statistiques.

9.5 Lignes directrices pour la diffusion de coefficients de variation

Avant de diffuser et/ou de publier toutes estimations établies à partir de l'EIJ, les utilisateurs devraient premièrement déterminer le niveau de qualité de cette estimation. Les niveaux de

qualité sont *acceptable*, *médiocre* et *inacceptable*. Les erreurs d'échantillonnage et non dues à l'échantillonnage, dont il a été question au chapitre 8.0, influencent la qualité des données. Aux fins du présent document, cependant, on ne déterminera le niveau de qualité d'une estimation qu'à partir d'une erreur d'échantillonnage dont rend compte le coefficient de variation indiqué à l'intérieur du tableau qui figure ci-dessous. Les utilisateurs devraient néanmoins s'assurer de lire le chapitre 8.0 pour être plus pleinement informés des caractéristiques relatives à la qualité de ces données.

On devrait premièrement déterminer le nombre de répondants retenus pour le calcul de l'estimation. Si ce nombre est inférieur à 30, il faudrait considérer l'estimation pondérée comme étant de qualité inacceptable.

Pour les estimations pondérées fondées sur les tailles d'échantillon de 30 ou plus, les utilisateurs devraient déterminer le coefficient de variation de l'estimation et suivre les lignes directrices relatives au niveau de qualité qui figurent ci-dessous. Celles-ci devraient être appliquées, pour la détermination du niveau de qualité d'une estimation, aux estimations pondérées arrondies.

On peut considérer qu'il est possible de divulguer toutes les estimations. Celles d'un niveau de qualité médiocre ou inacceptable doivent cependant être accompagnées d'une mise en garde pour avertir les utilisateurs subséquents.

Lignes directrices relatives au niveau de qualité de l'estimation

Niveau de qualité de l'estimation	Lignes directrices
1) Acceptable	<p>Les estimations proviennent d'une taille d'échantillon de 30 ou plus, et présentent de faibles coefficients de variation, de l'ordre de 0,0 à 16,5 %.</p> <p>Aucune mise en garde n'est requise.</p>
2) Médiocre	<p>Les estimations proviennent d'une taille d'échantillon de 30 ou plus, et présentent des coefficients de variation élevés, de l'ordre de 16,6 à 33,3 %.</p> <p>Ces estimations devraient être signalées par la lettre E (ou un quelconque identificateur similaire). Elles devraient être accompagnées d'une mise en garde avertissant les utilisateurs subséquents des niveaux élevés d'erreur associés aux estimations.</p>
3) Inacceptable	<p>Les estimations proviennent d'une taille d'échantillon inférieure à 30, ou présentent des coefficients de variation très élevés, supérieurs à 33,3 %.</p> <p>Statistique Canada recommande de ne pas diffuser d'estimations de qualité inacceptable. Si un utilisateur choisit cependant de le faire, ces estimations devraient alors être signalées à l'aide de la lettre F (ou d'un quelconque identificateur similaire) et devraient être accompagnées de la mise en garde suivante :</p> <p>« Nous informons l'utilisateur que ces estimations (désignées avec la lettre F) ne respectent pas les normes de qualité de Statistique Canada. Les conclusions qui reposeront sur ces données ne seront pas fiables et seront très probablement invalides. »</p>

9.6 Seuils pour la diffusion des estimations pour l'Enquête internationale auprès des jeunes

Le tableau ci-dessous fournit une indication de la précision des estimations démographiques ainsi que les seuils de diffusion associés aux trois niveaux de qualité de l'estimation présentés à la section précédente. Ces seuils proviennent des tables de coefficients de variation (CV) dont il sera question au chapitre 10.0.

Veillez noter que ces seuils de diffusion correspondent aux estimations de chiffres de population seulement. Dans le cas d'estimations de rapports, les utilisateurs ne devraient pas utiliser la valeur du numérateur (ni le dénominateur) afin de trouver le niveau de qualité de l'estimation correspondant. La règle 4 à la section 10.1 ainsi que l'exemple 4 à la section 10.1.1 expliquent la bonne procédure à suivre dans le cas d'un rapport.

Année d'études	Sexe	CV acceptable 0,0 à 16,5 %	CV médiocre 16,6 à 33,3 %	CV inacceptable > 33,3 %
7 ^e année	Masculin	900 et plus	250 à < 900	moins de 250
	Féminin	850 et plus	200 à < 850	moins de 200
	Tous	1 000 et plus	250 à < 1 000	moins de 250
8 ^e année	Masculin	1 100 et plus	300 à < 1 100	moins de 300
	Féminin	850 et plus	200 à < 850	moins de 200
	Tous	1 150 et plus	300 à < 1 150	moins de 300
9 ^e année	Masculin	1 400 et plus	400 à < 1 400	moins de 400
	Féminin	1 150 et plus	300 à < 1 150	moins de 300
	Tous	1 500 et plus	400 à < 1 500	moins de 400
Total	Masculin	1 150 et plus	300 à < 1 150	moins de 300
	Féminin	1 000 et plus	250 à < 1 000	moins de 250
	Tous	1 250 et plus	300 à < 1 250	moins de 300

10.0 Tables de variabilité d'échantillonnage approximative

Afin de fournir des coefficients de variation (CV) qui pourraient s'appliquer à une gamme étendue d'estimations catégoriques produites à partir de ce fichier de microdonnées et auxquels il serait facilement possible pour l'utilisateur d'avoir accès, un ensemble de tables de variabilité d'échantillonnage approximative a été produit. Ces tables de CV permettent à l'utilisateur d'obtenir un coefficient de variation approximatif fondé sur la taille de l'estimation calculée à partir des données de l'enquête.

Les coefficients de variation sont calculés à l'aide de la formule de la variance pour un échantillonnage aléatoire simple et en y incorporant un facteur qui reflète la nature du plan d'échantillonnage, qui est à plusieurs degrés et qui prévoit la formation de grappes. Ce facteur, appelé l'effet du plan, a été déterminé en calculant premièrement les effets du plan pour une gamme étendue de caractéristiques, puis en choisissant parmi ceux-ci une valeur modérée (habituellement le 75^e percentile) à utiliser à l'intérieur des tables de CV qui s'appliqueraient ensuite à l'ensemble entier des caractéristiques.

Le tableau ci-dessous indique la valeur modérée des effets du plan, ainsi que les tailles de l'échantillon et les chiffres de population selon l'année d'études qui ont été utilisés pour produire les tables de variabilité d'échantillonnage approximative de l'Enquête internationale auprès des jeunes (EIJ).

Année d'études	Sexe	Effet du plan	Taille de l'échantillon	Population
7 ^e année	Masculin	1,48	553	10 176
	Féminin	1,68	654	9 905
	Tous	1,73	1 207	20 081
8 ^e année	Masculin	1,67	550	10 741
	Féminin	1,55	598	9 549
	Tous	1,85	1 148	20 290
9 ^e année	Masculin	1,83	449	10 697
	Féminin	1,74	486	9 845
	Tous	1,98	935	20 541
Total	Masculin	1,63	1 552	31 613
	Féminin	1,69	1 738	29 299
	Tous	1,87	3 290	60 912

Tous les coefficients de variation inclus dans les tables de variabilité d'échantillonnage approximative sont approximatifs et donc non officiels. Des estimations de la variance réelle pour des variables précises peuvent être obtenues auprès de Statistique Canada, contre remboursement des frais. Étant donné que le CV approximatif est une estimation prudente, l'utilisation de la variance réelle estimée pourrait faire passer l'estimation d'un niveau de qualité à un autre. Par exemple, une estimation *médiocre* pourrait devenir *acceptable* si elle était fondée sur le calcul du CV exact.

Rappelez-vous que : Si le nombre d'observations sur lesquelles une estimation est basée est inférieur à 30, l'estimation pondérée est très probablement inacceptable et Statistique Canada recommande de ne pas diffuser une telle estimation, quelle que soit la valeur du coefficient de variation.

10.1 Comment utiliser les tables de coefficients de variation pour des estimations catégoriques

Les règles qui suivent devraient permettre à l'utilisateur de déterminer les coefficients de variation approximatifs à partir des tables de variabilité d'échantillonnage approximative pour des estimations du nombre, de la proportion ou du pourcentage de membres de la population visée par l'enquête possédant une certaine caractéristique et pour des rapports et des différences entre de telles estimations.

Règle 1 : Estimations du nombre de personnes possédant une caractéristique donnée (agrégats)

Le coefficient de variation dépend uniquement de la taille de l'estimation elle-même. Dans la table de variabilité d'échantillonnage approximative pour la région géographique appropriée, repérez le nombre estimé dans la colonne la plus à gauche (intitulée « Numérateur du pourcentage ») et suivez les astérisques (le cas échéant) jusqu'au premier chiffre rencontré. Ce chiffre est le coefficient de variation approximatif.

Règle 2 : Estimations de proportions ou de pourcentages de personnes possédant une caractéristique donnée

Le coefficient de variation d'une proportion estimée ou d'un pourcentage estimé dépend à la fois de la taille de la proportion ou du pourcentage et de la taille du total sur lequel la proportion ou le pourcentage repose. Les proportions estimées ou les pourcentages estimés sont relativement plus fiables que les estimations correspondantes du numérateur de la proportion ou du pourcentage, lorsque la proportion ou que le pourcentage repose sur un sous-groupe de la population. La proportion, par exemple, d'élèves qui aimaient beaucoup l'école en général est plus fiable que le nombre estimé d'élèves qui aimaient beaucoup l'école en général. (Remarquez que dans les tables la valeur des coefficients de variation diminue lorsqu'on les lit de gauche à droite.)

Lorsque la proportion ou que le pourcentage repose sur la population totale de la région géographique visée par la table, le CV de la proportion ou du pourcentage est le même que le CV du numérateur de la proportion ou du pourcentage. Dans ce cas, la règle 1 peut être appliquée.

Lorsque la proportion ou le pourcentage repose sur un sous-ensemble de la population totale (p. ex. comme ses membres d'un sexe ou d'une année d'études particulier(ère)) on devrait faire référence à la proportion ou au pourcentage (dans le haut de la table) et au numérateur de la proportion ou du pourcentage (dans la colonne de gauche de la table). L'intersection de la rangée et de la colonne appropriées donne le coefficient de variation.

Règle 3 : Estimations de différences entre des agrégats ou des pourcentages

L'erreur-type d'une différence entre deux estimations est approximativement égale à la racine carrée de la somme des carrés de chaque erreur-type considérée séparément. C'est-à-dire que l'erreur-type d'une différence ($\hat{d} = \hat{X}_1 - \hat{X}_2$) est :

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

où \hat{X}_1 est l'estimation 1, \hat{X}_2 est l'estimation 2 et α_1 et α_2 sont les coefficients de variation

de \hat{X}_1 et \hat{X}_2 respectivement. Le coefficient de variation de \hat{d} est donné par $\sigma_{\hat{d}}/\hat{d}$. Cette formule est exacte pour la différence entre des caractéristiques distinctes et non corrélées, mais n'est autrement qu'approximative.

Règle 4 : Estimations de rapports

Dans le cas où le numérateur est un sous-ensemble du dénominateur, le rapport devrait être converti en un pourcentage et la règle 2 appliquée. Cela s'appliquerait, par exemple, au cas où le dénominateur est le nombre d'élèves et le numérateur, le nombre d'élèves qui ont déjà bu de la bière, des panachés (coolers) ou du vin.

Dans le cas où le numérateur n'est pas un sous-ensemble du dénominateur, comme dans l'exemple du rapport du nombre d'élèves qui ont déjà bu de la bière, des panachés (coolers) ou du vin comparativement au nombre d'élèves qui ont déjà consommé une boisson très alcoolisée (gin, rhum, vodka, whisky) pure ou mélangée, l'erreur-type du rapport des estimations est approximativement égale à la racine carrée de la somme des carrés de chaque coefficient de variation considéré séparément multipliée par \hat{R} . C'est-à-dire que l'erreur-type d'un rapport ($\hat{R} = \hat{X}_1 / \hat{X}_2$) est :

$$\sigma_{\hat{R}} = \hat{R} \sqrt{\alpha_1^2 + \alpha_2^2}$$

où α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et de \hat{X}_2 respectivement. Le coefficient de variation de \hat{R} est donné par $\sigma_{\hat{R}}/\hat{R}$. La formule tendra à surestimer l'erreur si \hat{X}_1 et \hat{X}_2 sont corrélés positivement et à la sous-estimer si \hat{X}_1 et \hat{X}_2 sont corrélés négativement.

Règle 5 : Estimations de différences entre des rapports

Dans ce cas, les règles 3 et 4 sont combinées. On détermine premièrement les CV pour les deux rapports à l'aide de la règle 4, puis on trouve le CV de leur différence au moyen de la règle 3.

10.1.1 Exemples d'utilisation des tables de coefficients de variation pour des estimations catégoriques

Les exemples ci-dessous utilisent des données du fichier de l'EIJ et sont destinés à aider les utilisateurs à appliquer les règles que nous venons de présenter.

Exemple 1 : Estimations du nombre de personnes possédant une caractéristique donnée (agrégats)

Supposons qu'un utilisateur estime que 24 364 élèves de la 7^e à la 9^e année ont déjà bu de la bière, des panachés (coolers) ou du vin. Comment l'utilisateur détermine-t-il le coefficient de variation de cette estimation?

- 1) Reportez-vous à la table des coefficients de variation pour Toronto.
- 2) L'agrégat estimé (24 364) ne figure pas dans la colonne de gauche (la colonne « Numérateur du pourcentage »); il faut donc utiliser le chiffre qui s'en rapproche le plus, c'est-à-dire 25 000.

- 3) On trouve le coefficient de variation pour un agrégat estimé en se reportant à la première entrée autre que des astérisques sur cette rangée, c'est-à-dire 2,6 %.
- 4) Le coefficient de variation approximatif de l'estimation est donc 2,6 %. Le résultat selon lequel 24 364 (à être arrondi selon les lignes directrices pour l'arrondissement figurant à la section 9.1) élèves des écoles publiques et privées de la 7^e à la 9^e année avaient déjà bu de la bière, des panachés (coolers) ou du vin peut être publié sans réserve.

Exemple 2 : Estimations de proportions ou de pourcentages de personnes possédant une caractéristique donnée

Supposons qu'un utilisateur estime que $11\,663 / 29\,299 = 39,8\%$ des élèves de sexe féminin ont déjà bu de la bière, des panachés (coolers) ou du vin. Comment l'utilisateur détermine-t-il le coefficient de variation de cette estimation?

Dans cet exemple, l'utilisateur peut recourir à la table des coefficients de variation pour les élèves de sexe féminin et suivre la même méthode que dans l'exemple 1. La table des CV pour Toronto permet d'illustrer comment utiliser en même temps un pourcentage et une portion de numérateur.

- 1) Reportez-vous à la table des coefficients de variation pour Toronto.
- 2) Parce que l'estimation est un pourcentage fondé sur un sous-ensemble de la population totale (c'est-à-dire, les élèves de sexe féminin qui ont déjà bu de la bière, des panachés (coolers) ou du vin) il faut utiliser à la fois le pourcentage (39,8 %) et la portion numérateur du pourcentage (11 663) pour déterminer le coefficient de variation.
- 3) Le numérateur, 11 663, ne figure pas dans la colonne de gauche (la colonne « Numérateur du pourcentage »); il faut donc utiliser le chiffre qui s'en rapproche le plus, soit 12 500. De même, l'estimation du pourcentage ne figure dans l'en-tête d'aucune colonne; il faut donc utiliser la proportion qui s'en rapproche le plus, soit 40,0 %.
- 4) Le chiffre indiqué à l'intersection de la rangée et de la colonne utilisées, soit 4,0 %, est le coefficient de variation à employer.
- 5) Le coefficient de variation approximatif de l'estimation est donc 4,0 %. Le résultat selon lequel 39,8 % des élèves de sexe féminin ont déjà bu de la bière, des panachés (coolers) ou du vin peut être publié sans réserve.

Exemple 3 : Estimations de différences entre des agrégats ou des pourcentages

Supposons qu'un utilisateur estime que $14\,041 / 29\,299 = 47,9\%$ des élèves de sexe féminin ont fait du bénévolat dans les quatre dernières semaines contre $12\,658 / 31\,613 = 40,0\%$ des élèves de sexe masculin. Comment l'utilisateur détermine-t-il le coefficient de variation de la différence entre ces deux estimations?

- 1) L'utilisation de la table des coefficients de variation de Toronto de la même façon que celle décrite dans l'exemple 2, donne un CV de l'estimation pour les élèves de sexe féminin de 3,3 % et de 4,0 % pour les élèves de sexe masculin. .
- 2) En utilisant la règle 3, l'erreur-type d'une différence $(\hat{d} = \hat{X}_1 - \hat{X}_2)$ est :

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

où \hat{X}_1 est l'estimation 1 (les élèves de sexe féminin), \hat{X}_2 est l'estimation 2 (les élèves de sexe masculin) et α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et de \hat{X}_2 respectivement.

C'est-à-dire que l'erreur-type de la différence $\hat{d} = 0,479 - 0,400 = -0,079$ est :

$$\begin{aligned} \sigma_{\hat{d}} &= \sqrt{[(0,479)(0,033)]^2 + [(0,400)(0,040)]^2} \\ &= \sqrt{(0,000250) + (0,000256)} \\ &= 0,022 \end{aligned}$$

- 3) Le coefficient de variation de \hat{d} est donné par $\sigma_{\hat{d}} / \hat{d} = 0,022 / 0,079 = 0,278$.
- 4) Le coefficient de variation approximatif de la différence entre les estimations est donc 27,8 %. La différence entre les estimations est considérée inacceptable et Statistique Canada recommande de ne pas publier cette estimation. Cependant, si l'utilisateur choisit de publier cette donnée, elle devra être désignée ainsi en utilisant la lettre E (ou un autre identificateur semblable) et être accompagnée d'un avertissement mettant les utilisateurs subséquents en garde contre les hauts taux d'erreur associés à l'estimation.

Exemple 4 : Estimations de rapports

Supposons qu'un utilisateur estime que 2 749 des élèves de sexe féminin et 3 023 des élèves de sexe masculin ayant déjà bu de la bière, des panachés (coolers) ou du vin ont déclaré en avoir consommé la première fois quand ils avaient moins de 10 ans. L'utilisateur est intéressé à comparer l'estimation des élèves de sexe féminin à celle des élèves de sexe masculin sous la forme d'un rapport. Comment l'utilisateur détermine-t-il le coefficient de variation de cette estimation?

- 1) Tout d'abord, cette estimation est une estimation d'un rapport, où le numérateur de l'estimation (\hat{X}_1) est le nombre d'élèves de sexe féminin ayant moins de 10 ans lorsqu'elles ont bu de la bière, des panachés (coolers) ou du vin la première fois. Le dénominateur (\hat{X}_2) est le nombre d'élèves de sexe masculin ayant moins de 10 ans quand ils ont bu de la bière la première fois.
- 2) Pour les élèves de sexe féminin, reportez-vous à la table des coefficients de variation pour les élèves de sexe féminin. Pour les élèves de sexe masculin, reportez-vous à la table des coefficients de variation pour les élèves de sexe masculin.
- 3) Le numérateur de cette estimation de rapport est 2 749. Le chiffre qui s'en rapproche le plus dans la table des coefficients de variation pour les élèves de sexe féminin est 2 500. On trouve le coefficient de variation pour cette estimation en se reportant à la première entrée autre que des astérisques sur cette rangée, soit 9,8 %.
- 4) Le dénominateur de cette estimation de rapport est 3 023. Le chiffre qui s'en rapproche le plus dans la table des coefficients de variation pour les élèves de sexe

masculin est 3 000. On trouve le coefficient de variation pour cette estimation en se reportant à la première entrée autre que des astérisques sur cette rangée, soit 9,7 %.

- 5) Le coefficient de variation approximatif de l'estimation du rapport est donc donné par la règle 4, qui est :

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^2 + \alpha_2^2}$$

où α_1 et α_2 sont les coefficients de variation de \hat{X}_1 et \hat{X}_2 respectivement.

C'est-à-dire que :

$$\begin{aligned} \alpha_{\hat{R}} &= \sqrt{(0,098)^2 + (0,097)^2} \\ &= \sqrt{0,0096 + 0,0094} \\ &= 0,138 \end{aligned}$$

- 6) Le rapport obtenu entre les élèves de sexe féminin et les élèves de sexe masculin ayant bu de la bière, des panachés (coolers) ou du vin la première fois lorsqu'ils avaient moins de 10 ans est 2 749 / 3 023, c'est-à-dire 0,91 (à être arrondi selon les lignes directrices pour l'arrondissement figurant à la section 9.1). Le coefficient de variation de cette estimation est 13,8 %, ce qui fait qu'on peut le diffuser sans réserve.

Exemple 5 : Estimations de différences de rapports

Supposons que l'utilisateur estime que le rapport entre les élèves de sexe féminin et les élèves de sexe masculin qui ont déjà bu de la bière, des panachés (coolers) ou du vin est de 1,050 pour la 8^e année et de 0,919 pour la 9^e année. L'utilisateur est intéressé à comparer les deux rapports pour voir s'il y a une différence statistique entre ceux-ci. Comment l'utilisateur détermine-t-il le coefficient de variation de cette différence?

- 1) Tout d'abord, calculez le coefficient de variation approximatif pour le rapport de la 8^e année (\hat{R}_1) en utilisant la table des CV pour la 8^e année, tous les élèves et le rapport de la 9^e année (\hat{R}_2) en utilisant la table des CV pour la 9^e année, tous les élèves tels qu'il est décrit dans l'exemple 4. Le CV approximatif pour le rapport de la 8^e année est 8,7 % et 8,8 % pour celui de la 9^e année.
- 2) En utilisant la règle 3, l'erreur-type d'une différence ($\hat{d} = \hat{R}_1 - \hat{R}_2$) est :

$$\sigma_{\hat{d}} = \sqrt{(\hat{R}_1 \alpha_1)^2 + (\hat{R}_2 \alpha_2)^2}$$

où α_1 et α_2 sont les coefficients de variation de \hat{R}_1 et \hat{R}_2 respectivement. C'est-à-dire que l'erreur-type de la différence $\hat{d} = 1,050 - 0,919 = 0,131$ est :

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(1,0501)(0,0870)]^2 + [(0,9187)(0,0884)]^2} \\ &= \sqrt{(0,0083) + (0,0066)} \\ &= 0,122\end{aligned}$$

3) Le coefficient de variation de \hat{d} est donné par $\sigma_{\hat{d}} / \hat{d} = 0,122 / (0,131) = 0,931$

4) Le coefficient de variation approximatif de la différence entre les estimations est donc 93,1 %. La différence entre les estimations est considérée inacceptable et Statistique Canada recommande de ne pas la publier. Cependant, si l'utilisateur choisit de publier cette donnée, elle devra être désignée ainsi en utilisant la lettre F (ou un autre identificateur semblable) et être accompagnée d'un avertissement mettant les utilisateurs subséquents en garde contre les hauts taux d'erreur associés à l'estimation.

10.2 Comment utiliser les tables de coefficients de variation pour obtenir des limites de confiance

Bien que les coefficients de variation soient beaucoup utilisés, l'intervalle de confiance d'une estimation est une mesure plus intuitivement significative de l'erreur d'échantillonnage. Un intervalle de confiance constitue une déclaration du niveau de confiance selon laquelle la valeur vraie pour la population se situe à l'intérieur d'une gamme précisée de valeurs. Par exemple, un intervalle de confiance de 95 % peut être décrit comme suit :

Si l'échantillonnage de la population est répété indéfiniment, chaque échantillon menant à un nouvel intervalle de confiance pour une estimation, l'intervalle englobera alors dans 95 % des échantillons la valeur vraie de la population.

En utilisant l'erreur-type d'une estimation, des intervalles de confiance pour des estimations peuvent être obtenues en partant de l'hypothèse qu'aux termes d'un échantillonnage répété de la population, les diverses estimations obtenues pour une caractéristique donnée de la population se répartiront normalement autour de la valeur vraie de la population. Selon cette hypothèse, il y a environ 68 chances sur 100 que l'écart entre une estimation de l'échantillon et la valeur vraie pour la population soit inférieur à une erreur-type, environ 95 chances sur 100 que l'écart soit inférieur à deux erreurs-types et environ 99 chances sur 100 que l'écart soit inférieur à trois erreurs-types. Ces différents degrés de confiance sont désignés sous le nom de niveaux de confiance.

Des intervalles de confiance pour une estimation \hat{X} sont généralement exprimés sous forme de deux chiffres, un inférieur et un supérieur à l'estimation, comme étant $(\hat{X} - k, \hat{X} + k)$, où k est déterminé suivant le niveau de confiance désiré et l'erreur d'échantillonnage de l'estimation.

Des intervalles de confiance pour une estimation peuvent être calculés directement à partir des tables de variabilité d'échantillonnage approximative, en déterminant d'abord à partir de la table appropriée le coefficient de variation de l'estimation \hat{X} , puis en utilisant la formule suivante pour le convertir à un intervalle de confiance ($IC_{\hat{x}}$) :

$$IC_{\hat{x}} = (\hat{X} - t\hat{X}\alpha_{\hat{x}}, \hat{X} + t\hat{X}\alpha_{\hat{x}})$$

où $\alpha_{\hat{x}}$ est le coefficient de variation déterminé de \hat{X} , et

- $t = 1$ si l'on désire un intervalle de confiance de 68 %;
- $t = 1,6$ si l'on désire un intervalle de confiance de 90 %;
- $t = 2$ si l'on désire un intervalle de confiance de 95 %;
- $t = 2,6$ si l'on désire un intervalle de confiance de 99 %.

Nota : Les lignes directrices pour la diffusion des estimations s'appliquent également aux intervalles de confiance. S'il est impossible, par exemple, de diffuser une estimation, on ne peut alors pas non plus communiquer un intervalle de confiance.

10.2.1 Exemple d'utilisation des tables de coefficients de variation pour obtenir des limites de confiance

Un intervalle de confiance de 95 % pour la proportion estimée des élèves de sexe féminin ayant déjà bu de la bière, des panachés (coolers) ou du vin (d'après l'exemple 2 à la section 10.1.1) serait calculé comme suit :

$$\hat{X} = 39,8 \% \text{ (ou exprimé sous forme de proportion } 0,398)$$

$$t = 2$$

$$\alpha_{\hat{x}} = 4,0 \% \text{ (0,040 exprimé sous forme de proportion) est le coefficient de variation de cette estimation, tel que déterminé à partir des tables.}$$

$$IC_{\hat{x}} = \{0,398 - (2) (0,398) (0,040), 0,398 + (2) (0,398) (0,040)\}$$

$$IC_{\hat{x}} = \{0,398 - 0,032, 0,398 + 0,032\}$$

$$IC_{\hat{x}} = \{0,366, 0,430\}$$

Avec un intervalle de confiance de 95 %, on peut dire qu'entre 36,6 % et 43,0 % des élèves de sexe féminin ont déjà bu de la bière, des panachés (coolers) ou du vin.

10.3 Comment utiliser les tables de coefficients de variation pour effectuer un test t

Des erreurs-types peuvent aussi être utilisés pour effectuer des tests d'hypothèses, une procédure destinée à distinguer des paramètres d'une population à l'aide d'estimations d'un échantillon. Ces estimations peuvent être des chiffres, des moyennes, des pourcentages, des rapports, etc. Les tests peuvent être effectués à divers niveaux de signification, où un niveau de signification est la probabilité de conclure que les caractéristiques sont différentes quand, en fait, elle sont identiques.

Supposons que \hat{X}_1 et \hat{X}_2 sont des estimations d'un échantillon pour deux caractéristiques qui nous intéressent. Supposons également que l'erreur-type de la différence $\hat{X}_1 - \hat{X}_2$ est $\sigma_{\hat{d}}$.

Si $t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{a}}}$ se situe entre -2 et 2, aucune conclusion à propos de la différence entre les caractéristiques n'est alors justifiée au niveau de signification de 5 %. Si, cependant, ce rapport est inférieur à -2 ou supérieur à +2, la différence observée est significative au niveau de 0,05. C'est-à-dire que la différence entre les estimations est significative.

10.3.1 Exemple d'utilisation des tables de coefficients de variation pour effectuer un test t

Supposons que l'utilisateur désire tester, au niveau de signification de 5 %, l'hypothèse selon laquelle il n'y a pas de différence entre la proportion des élèves de sexe féminin et des élèves de sexe masculin ayant déclaré avoir fait du bénévolat au cours des quatre dernières semaines. D'après l'exemple 3 à la section 10.1.1, il s'est avéré que l'erreur-type de la différence entre ces deux estimations était 0,022. Par conséquent,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{a}}} = \frac{0,479 - 0,400}{0,022} = \frac{0,079}{0,022} = 3,59$$

Puisque $t = 3,59$ est supérieur à 2, il faut en conclure qu'il existe une différence significative entre les deux estimations au niveau de signification de 0,05.

10.4 Coefficients de variation pour des estimations quantitatives

Il faudrait produire des tables spéciales afin de déterminer l'erreur d'échantillonnage d'estimations quantitatives, ce qui n'a pas été fait, parce que la plupart des variables pour l'EIJ sont principalement de nature catégorique.

En général cependant, le coefficient de variation d'un total quantitatif sera supérieur au coefficient de variation de l'estimation de la catégorie correspondante (c'est-à-dire l'estimation du nombre de personnes retenues dans l'estimation quantitative). S'il est impossible de diffuser l'estimation de la catégorie correspondante, on ne pourra pas non plus communiquer l'estimation quantitative. Par exemple, le coefficient de variation du nombre total d'amis qui ont volé quelque chose dans un magasin serait supérieur au coefficient de variation de la proportion correspondante d'élèves comptant un ou plusieurs amis qui ont volé quelque chose dans un magasin. Si, par conséquent, le coefficient de variation de la proportion est inacceptable (rendant la proportion non diffusable), il en sera de même du coefficient de variation de l'estimation quantitative correspondante (rendant cette estimation quantitative non diffusable).

Des coefficients de variation de telles estimations peuvent être calculés, au besoin, pour une estimation précise à l'aide d'une technique appelée pseudo-répétition, ce qui veut dire diviser les enregistrements inclus dans les fichiers de microdonnées en sous-groupes (ou répétitions) et déterminer la variation à l'intérieur de l'estimation de répétition en répétition. Les utilisateurs qui désirent calculer des coefficients de variation pour des estimations quantitatives peuvent communiquer avec Statistique Canada afin d'obtenir des conseils sur l'allocation d'enregistrements à des répétitions appropriées et sur les formules à employer à l'intérieur de ces calculs.

10.5 Tables des coefficients de variation

Consulter le fichier EIJ2006_CVTabF.pdf pour les tables de coefficients de variation.

11.0 Pondération

Les poids statistiques associés à chaque enregistrement représentent le nombre d'élèves sélectionnés que contient chacun des enregistrements. La pondération des données de l'Enquête internationale auprès des jeunes comprend plusieurs étapes qui sont décrites dans les paragraphes suivants :

1. Pondération initiale de l'échantillon (le poids de l'école)

La première étape est de calculer le poids initial ($Poids_1$) pour chaque unité sélectionnée (école-année d'études). Pour une unité donnée, ceci correspond à l'inverse de la probabilité de sélection dans la strate concernée. Cette probabilité est proportionnelle au nombre d'élèves de l'année d'études donnée de l'école. Comme il s'agit à ce niveau d'un échantillonnage avec probabilité proportionnelle à la taille, il a fallu mettre un certain nombre de grandes écoles dans des strates distinctes à tirage complet. Ces écoles ont reçu une valeur initiale de pondération correspondant à l'unité.

2. Élimination des écoles inadmissibles

Pendant la collecte des données, on a jugé que cinq écoles étaient inadmissibles (n'ayant pas l'année d'études pour laquelle elles avaient été sélectionnées). Ces écoles ont été éliminées. Il n'y a pas eu d'ajustement de poids pour les écoles admissibles, d'où

$$Poids_2 = Poids_1$$

3. Ajustement pour la non-réponse au niveau de l'école

On a observé de la non-réponse au sein des unités école-année d'études initialement sélectionnées. Plusieurs facteurs peuvent expliquer la non-réponse au niveau de l'école tels le refus de participer des écoles ou l'incapacité de compléter les entrevues durant la période prévue pour la collecte des données.

Le calcul de cet ajustement de non-réponse au niveau des écoles a varié selon que l'unité école-année d'études appartenait à une strate à tirage partiel ou à tirage complet, et ce, parce que le poids déterminé par le plan de sondage tient compte de la taille des écoles dans les strates à tirage partiel, mais devient de l'égal (sans égard à la taille des écoles) dans le cas des strates à tirage complet.

- **Unités des strates à tirage partiel**

Pour les unités de la strate h pour l'année d'études g , l'ajustement se définit ainsi :

$$ajus_{\text{école}_{nr}} = \frac{\text{Nombre d'écoles admissibles de la strate } h \text{ pour l'année d'études } g}{\text{Nombre d'écoles répondantes de la strate } h \text{ pour l'année d'études } g}$$

- **Unités des strates à tirage complet**

Pour les unités de la strate h pour l'année d'études g , l'ajustement se définit ainsi :

$$ajus_{\text{école}_{nr}} = \frac{\sum Poids_2 \text{ pour les écoles répondantes} + \sum Poids_2 \text{ pour les écoles non répondantes}}{\sum Poids_2 \text{ pour les écoles répondantes}}$$

À la suite de cette étape, la pondération est $Poids_3 = Poids_2 * ajus_{\text{école}_{nr}}$

4. Ajustement pour la sélection d'une classe (le poids de la classe)

Cet ajustement correspond au deuxième degré d'échantillonnage lorsqu'une classe est sélectionnée de façon aléatoire parmi toutes celles de la même année d'études dans l'école choisie. Comme on a sélectionné une seule classe par école-année d'études, l'ajustement se fait en multipliant le poids obtenu à l'étape précédente par le nombre total de classes de l'école pour cette année d'études. On retrouve ce nombre sur la formule de sélection de classes.

$$Poids_4 (\text{poids de la classe}) = Poids_3 * \text{nombre de classes}$$

5. Ajustement pour la non-réponse de classe

Cet ajustement correspond à la non-réponse au niveau de la classe. La non-réponse au niveau de la classe est définie comme étant tous cas dont on connaît le nombre de classes (et qui est positif) mais pour lesquelles aucun élève n'a répondu. Le facteur de ajustement est défini comme suit :

$$ajus_{\text{classe_nr}} = \frac{\sum Poids_4 \text{ pour les classes répondantes} + \sum Poids_4 \text{ pour les classes non répondantes}}{\sum Poids_4 \text{ pour les classes répondantes}}$$

Le poids résultant pour cette étape est $Poids_5 = Poids_4 * ajus_{\text{classe_nr}}$.

Nota : Puisque tous les élèves de la classe sélectionnée ont été recensés, cette étape donne également le poids des élèves.

6. Ajustement pour la non-réponse des élèves

Cet ajustement a pour objet de compenser pour la non-réponse au niveau des élèves. Les raisons principales de ce genre de non-réponse sont : une autorisation parentale non obtenue, le refus d'un élève de participer ou l'absence de l'élève le jour de la collecte en classe. L'ajustement consiste à multiplier le poids obtenu à l'étape précédente par le ratio suivant :

$$ajus_{\text{élève_nr}} = \frac{\text{Nombre d'élèves admissibles dans la classe sélectionnée}}{\text{Nombre d'élèves répondants dans la classe sélectionnée}}$$

On obtient ainsi $Poids_6 = Poids_5 * ajus_{\text{élève_nr}}$.

7. Ajustement de post-stratification

On apporte un ajustement de poids d'échantillonnage pour les élèves des écoles publiques, de sorte qu'il y ait concordance avec les chiffres de dénombrement des inscriptions dans certains groupes (post-strates). Ces chiffres ont été fournis par le Conseil scolaire du district de Toronto (CSDT) à la fin de la collecte des données pour les post-strates année d'études-sexe. Le ratio entre le nombre réel d'élèves dans une post-strate et le nombre estimé par le plan d'échantillonnage représente cet ajustement. Pour les unités appartenant à la post-strate p , l'ajustement de la post-stratification est défini comme suit :

$$ajus_{\text{post_strate}} = \frac{\text{Totaux de dénombrement de la post - strate}_p}{\sum Poids_5 \text{ pour les enregistrements de la post - strate}_p}$$

Pour les écoles privées, on n'a pas obtenu de tels chiffres de dénombrement et, par conséquent, le facteur d'ajustement est simplement l'unité.

Le poids final d'échantillonnage de chaque enregistrement est le produit d'ajustement de poids des élèves multiplié par l' $ajus_{post_strate}$, c.-à-d. $WTPP = Poids_6 * ajus_{post_strate}$.

8. Mesure supplémentaire de prévention de divulgation

Pour empêcher toute divulgation de données relatives aux écoles dans le cas du fichier de microdonnées à grande diffusion (FMGD), nous avons perturbé aléatoirement le poids d'un certain nombre d'enregistrements (moins de 2 %). L'incidence de cette mesure est négligeable sur la répartition des valeurs de poids selon l'année d'études et le sexe.

12.0 Questionnaire

Le questionnaire de l'Enquête internationale auprès des jeunes (EIJ) a servi à recueillir des données pour le cas du volet canadien de l'étude en 2006. Le fichier EIJ2006_QuestF.pdf renferme le questionnaire français.

13.0 Cliché d'enregistrement à valeurs univariées

Consulter le fichier EIJ2006_LvCds.pdf pour le cliché d'enregistrement à chiffres univariés.