



Statistique Canada

www.statcan.gc.ca

Le respect de la confidentialité et les Centres de données de recherche

Jean-Louis Tambay et Denis Gonthier

20 avril 2012



Statistique
Canada

Statistics
Canada

Canada



Plan de la présentation

- Concepts et définitions
- Contrôle de la divulgation aux Centres de données de recherche
- Règles de confidentialité et catégories d'enquêtes



Concepts et définitions



Confidentialité et divulgation

- Statistique Canada protège la confidentialité de l'identité et des valeurs des répondants lors de la diffusion de données.
- La divulgation vise le rattachement inopportun de renseignements à une unité de données, qu'il s'agisse d'une personne ou d'un organisme.



Types de divulgation

- Divulgation de l'identité
 - Se produit lorsqu'un sujet (le répondant) peut être identifié dans des données diffusées
 - P.ex., identification sur un fichier de microdonnées
- Divulgation d'attributs
 - Se produit lorsque des renseignements confidentiels sont divulgués et qu'ils peuvent être associés à une unité (personne, organisme...)
 - P.ex., la seule cellule non vide dans une rangée
- Divulgation par inférence
 - Se produit lorsque des informations concernant une unité peuvent être déduites avec un niveau de certitude élevé (faible niveau d'incertitude)



Divulcation résiduelle

- Se produit lorsque des renseignements sont divulgués par l'association de renseignements diffusés avec d'autres renseignements accessibles
- Exemples :
 - Reconstitution de cellules qui ont été supprimées dans un tableau
 - Obtention de renseignements concernant de petites régions à partir de données pour des régions géographiques qui se recoupent
 - Demandes ciblées pour déjouer les mesures de protection d'un système d'interrogation en direct



Le risque de divulgation varie selon les caractéristiques des données

- Données d'échantillon ou de recensement
- Données d'enquête ou données administratives
- Données sociales ou économiques
- Genre de variables
- Genre de sorties

Plan de sondage – échantillon ou recensement

- Le risque de divulgation est plus élevé lorsqu'il s'agit de données de recensement
 - Peu de doute si une personne unique est identifiée
 - Peu/pas d'incertitude lors d'une divulgation d'attributs
- L'échantillonnage réduit le risque de divulgation
 - Les personnes uniques dans l'échantillon ne sont pas forcément uniques dans la population
 - La divulgation d'attributs doit tenir compte de la variabilité échantillonnale
 - Le plan de sondage peut avoir un effet sur le risque
 - On peut utiliser l'échantillonnage pour réduire le risque



Source de données – données d'enquête ou administratives

- Les données administratives présentent souvent les mêmes genres de risques que les données de recensement...
- ... et les fournisseurs des données peuvent diffuser eux-même des informations, ce qui augmente le risque de divulgation résiduelle
- Certains de nos fichiers d'enquêtes ont les deux genres de données



Données sociales ou économiques

- Données d'enquêtes économiques
 - Petites populations par domaine (industrie/région)
 - Utilisation de strates à tirage complet (100 %)
 - Données quantitatives à distribution asymétrique
 - Sorties en tableaux – problème de prédominance
- Données d'enquêtes sociales ou de ménages
 - Grandes populations, faibles taux de sondage
 - Plus « riches » en données, structure hiérarchique
 - Plusieurs moyens de diffusion de données, y compris sous forme de microdonnées anonymisées

Genres de variables présentes sur les microdonnées

- Identifiants directs
 - Noms, adresses, nos. téléphone, NAS, ...
 - Souvent absents de fichiers maîtres
- Identifiants indirects
 - Peuvent aider à l'identification (code postal, âge, état matrimonial, langue, lieu de naissance, occupation...)
 - Présentent parfois des risques pour la divulgation
- Variables dites « sensibles »
 - Caractéristiques qui se rapportent à la vie privée ou à l'entreprise du répondant, qui ne sont habituellement pas connues



Genres de sorties

- Fichiers de microdonnées anonymisées
 - Les microdonnées présentent le plus grand risque
 - Doivent subir un traitement intensif pour assurer la confidentialité – incidence sur la qualité des données
- Tableaux de fréquence ou quantitatifs
 - Risques réduits – surtout pour les enquêtes-ménages
 - Possibilité de divulgation d'identité ou d'attributs
- Résultats analytiques
 - Présentent habituellement le moindre risque
 - Des sorties « standard » de programmes peuvent présenter des risques (p.ex., les valeurs extrêmes)

Exemple – Tableau de fréquence

Âge	Groupe A	Groupe B	Groupe C	Total
<20	0	17	0	17
20-40	54	36	12	102
40-60	33	0	45	78
60-80	9	0	13	22
80+	0	6	1	7
Total	96	59	71	226

Modes de contrôle de la divulgation

- Méthodes d'accès restreint aux données
 - Centres de données de recherche, accès à distance, contrats de licence...
- Méthodes de diffusion restreinte des données
 - Méthodes réduisant l'information : suppression de données, regroupement de catégories, arrondissement*, troncature par le haut...
 - Méthodes de perturbation des données : ajout de bruit, permutation de données, imputation...

* - choix du programme d'accès à distance en temps réel



Contrôle de la divulgation aux Centres de données de recherche



Contrôle de la divulgation aux Centres de données de recherche

- Les données d'enquêtes-ménages présentent généralement moins de risque – mais la multitude du nombre et des genres de sorties augmente le risque de divulgation résiduelle
- Quelques préoccupations
 - Divulgence d'attributs pour des caractéristiques qui seraient identifiantes ou de nature délicate
 - Diffusion d'informations additionnelles qui pourraient être rattachées à des répondants sur les fichiers de microdonnées (en utilisant leurs poids d'enquête) – ce qui contribuerait à augmenter leur risque d'identification



Ne pas diffuser

- Résultats analytiques qui donnent des valeurs individuelles
 - Nuages de points, valeurs extrêmes, résidus
 - Certaines valeurs ont été supprimées/perturbées sur les fichiers de microdonnées anonymisées
- Résultats pour des petits ensembles
 - Cellules de tableaux ayant peu de répondants
 - Totaux et statistiques basés sur peu d'observations
- Divulgence d'attributs présentant des risques
 - Les tableaux qui comportent des cellules complètes (une seule cellule non vide dans une rangée/colonne) doivent faire objet d'une attention particulière



À surveiller

- Divulcation résiduelle
 - Tableaux liés permettant de reproduire un tableau à risque (p.ex., tableaux pour l'ensemble de la population et pour la population non autochtone)
 - Résultats pour des sous-ensembles identiques à quelques observations près
- Résultats analytiques qui équivalent à des statistiques simples ou des tableaux
 - Si c'est le cas, s'assurer que les statistiques ou tableaux sont diffusables
- Résultats à caractère géographique

Résultats analytiques à surveiller

- Centiles qui se rapportent à des valeurs extrêmes (p.ex., 90^e centile de 20 observations)
- Histogrammes (contrôler le tableau de fréquence)
- Modèles saturés dont les covariables sont tous dichotomiques (reproduire et contrôler le tableau)
- Covariances et corrélations avec une variable x_i dichotomique (donne des sous-totaux pour y_i)

$$\hat{\sigma}_{xy} = \frac{\sum_i w_i x_i y_i - (\sum_i w_i x_i)(\sum_i w_i y_i) / (\sum_i w_i)}{(\sum_i w_i)}$$

Exemple de sortie – matrice de corrélations

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 /
N = 2221 / **WEIGHT Var = WT**

	DAGEGRP1	DAGEGRP2	DAGEGRP3	MALE	OWNER	RESTRICT
DAGEGRP1	1.00000	-0.37498	-0.37953	0.01260	-0.12112	-0.09834
	0.0	0.0001	0.0001	0.5529	0.0001	0.0001
DAGEGRP2	-0.37498	1.00000	-0.48442	0.00896	-0.14627	-0.07146
	0.0001	0.0	0.0001	0.6730	0.0001	0.0008
DAGEGRP3	-0.37953	-0.48442	1.00000	0.00912	0.20420	0.02046
	0.0001	0.0001	0.0	0.6675	0.0001	0.3352
MALE	0.01260	0.00896	0.00912	1.00000	0.02423	-0.02648
	0.5529	0.6730	0.6675	0.0	0.2537	0.2122
OWNER	-0.12112	-0.14627	0.20420	0.02423	1.00000	-0.03539
	0.0001	0.0001	0.0001	0.2537	0.0	0.0954
RESTRICT	-0.09834	-0.07146	0.02046	-0.02648	-0.03539	1.00000
	0.0001	0.0008	0.3352	0.2122	0.0954	0.0

Exemple de sortie – Proc Univariate

Univariate Procedure

Variable=INCOME

Weight= WT

Moments

N	2133	Sum Wgts	2809775
Mean	45009.73	Sum	1.265E11
Std Dev	909932.7	Variance	8.28E11
Skewness	.	Kurtosis	.
USS	7.458E15	CSS	1.765E15
CV	2021.636	Std Mean	542.8421
T:Mean=0	82.91495	Pr> T	0.0001
Num ^= 0	2109	Num > 0	2109
M(Sign)	1054.5	Pr>= M	0.0001
Sgn Rank	1112498	Pr>= S	0.0001

Exemple de sortie – Proc Univariate

Quantiles (Def=5)

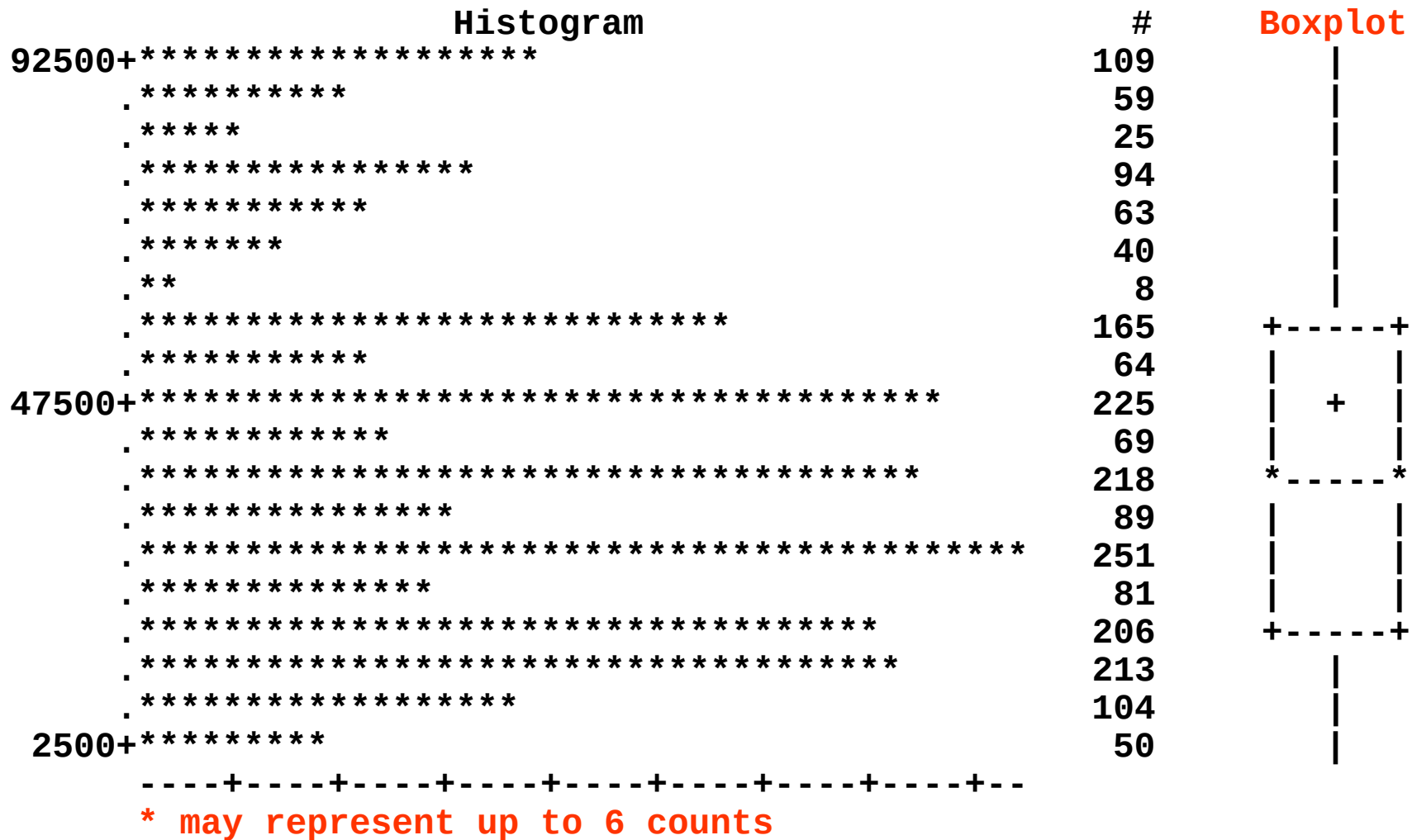
100% Max	93782.16	99%	92968.52
75% Q3	56340.27	95%	90231.07
50% Med	37135.4	90%	78679.8
25% Q1	19572.83	10%	12628.7
0% Min	0	5%	8731.114
		1%	0

Range	93782.16
Q3 - Q1	36767.45
Mode	0

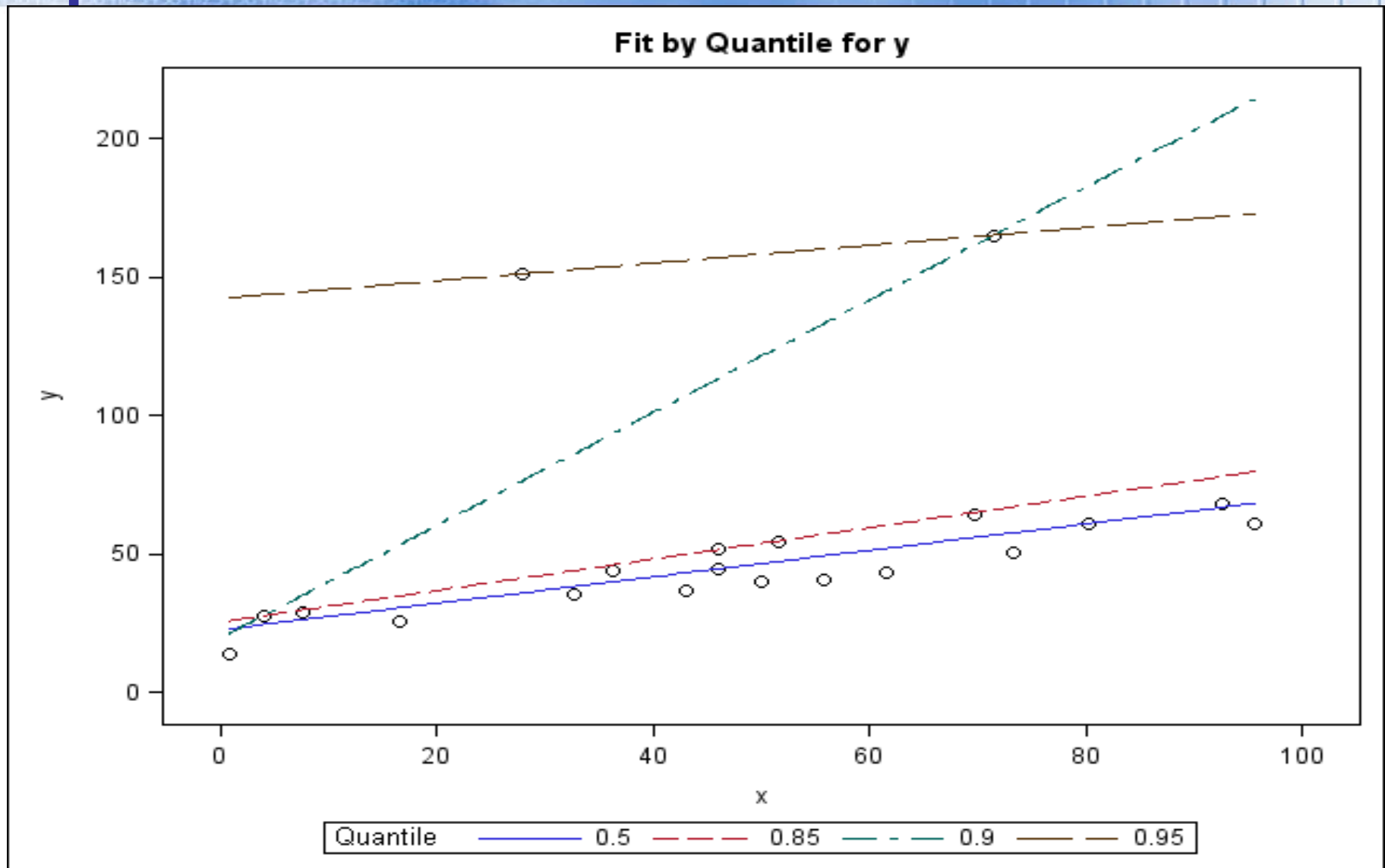
Extremes

Lowest	Obs	Highest	Obs
0 (	2150)	93552.71 (	1319)
0 (	2025)	93597.12 (	572)
0 (	1828)	93632.31 (	534)
0 (	1797)	93681.39 (	226)
0 (	1769)	93782.16 (	62)

Exemple de sortie – histogramme et tracé en rectangle et moustaches



Exemple de sortie – les régressions quantiles relient les observations





Résultats à caractère géographique

- Les variables géographiques augmentent le risque d'identification pour les microdonnées – ils y sont présents sous forme agrégée seulement
- Les résultats géographiques détaillés portent souvent sur des petits ensembles de données
- Ils peuvent dévoiler la location d'unités de l'échantillon – ce qui peut augmenter le risque d'identification pour certaines unités
- C'est pourquoi la diffusion de résultats géographiques est scrutée plus profondément



Règles de confidentialité et catégories d'enquêtes



Règles de contrôle de confidentialité

Catégories de règles pour les enquêtes dans les CDR

- Règles principales
- Variantes relatives à l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et à un effectif minimum de 10
- Règles pour le recensement
- Règles de l'Enquête sur le milieu de travail et les employés
- Règles de l'Enquête longitudinale sur les immigrants au Canada
- Règles applicables aux enquêtes post-censitaires



Règles principales

S'appliquent à la majorité des enquêtes, par exemple :

- Enquête sur les dépenses des ménages
- Enquête sur la dynamique du travail et du revenu
- Enquête auprès des jeunes en transition
- Enquête longitudinale nationale sur les enfants et les jeunes
- Enquête nationale auprès des diplômés
- Enquête nationale sur la santé de la population
- Enquête sur la population active
- Enquête sur la sécurité financière
- Enquête sociale générale

Règles principales

1) Quelle est la fréquence non pondérée requise au minimum?	5
2) Est-ce permis de diffuser des résultats descriptifs non pondérés?	OUI, fréquence minimale requise de 15 (Note 1)
3) Peut-on diffuser à la fois des résultats descriptifs pondérés et non pondérés pour cette enquête?	NON (Note 2 - exception possible)
4) Est-il requis d'arrondir tous les résultats descriptifs pondérés et, si oui, quelle est la base d'arrondissement?	NON (Note 3)
5) Peut-on diffuser les résultats d'un modèle avec et sans pondération dans le cas de cette enquête?	OUI
6) Critères géographiques d'arrondissement : La règle suivante s'applique-t-elle? Les fréquences au niveau infraprovincial doivent respecter le nombre minimal d'observations requises dans une cellule et être pondérées et elles doivent faire l'objet d'un arrondissement à la base 50.	OUI

Variantes pour l'ESCC et un effectif minimum de 10

L'ESCC utilise les règles principales avec une différence quant à la géographie (certains échantillons étant plus grands)

ESCC et critères géographiques d'arrondissement

- L'arrondissement n'est pas nécessaire pour les questions de « contenu commun » dans l'ESCC 1.1, 2.1, 3.1 et 2007
- L'arrondissement s'applique au contenu en sous-échantillon (optionnel) des enquêtes de la série.1, de toutes les enquêtes de la série.2 et de l'Enquête sur l'accès aux services de santé.
- Quelques enquêtes utilisent les règles principales mais ont un effectif minimum de 10:
 - Enquête canadienne sur les mesures de la santé
 - Enquête nationale sur le travail et la santé du personnel infirmier

Règles pour le recensement

1) Quelle est la fréquence non pondérée requise au minimum?	4 (valeur pondérée minimale de 10)	Valeur minimale de 4 adoptée à la Division des opérations du recensement pour leurs tableaux spéciaux. Il s'agit là d'une norme publique.
2) Est-ce permis de diffuser des résultats descriptifs non pondérés?	NON	Permission nécessaire pour description de base de l'échantillon (justification). Les fréquences non pondérées doivent être arrondies au multiple de 5.
3) Peut-on diffuser à la fois des résultats descriptifs pondérés et non pondérés pour cette enquête?	NON	Non, sauf pour la description de l'échantillon comme il est indiqué à la ligne précédente.
4) Est-il requis d'arrondir tous les résultats descriptifs pondérés et, si oui, quelle est la base d'arrondissement?	OUI, base 5	On peut procéder par arrondissement simple ou aléatoire. Les fréquences de moins de 10 sont arrondies à la base 10.

Règles pour le recensement (suite)

5) Peut-on diffuser les résultats d'un modèle avec et sans pondération dans le cas de cette enquête?	NON	Si on diffuse des données non pondérées, il n'y a pas de données pondérées qui puissent l'être pour cette demande en particulier.
6) Critères géographiques : La population de l'entité géographique pour laquelle des résultats sont générés doit dépasser un certain seuil.	Population totale de la région est au minimum de 250 personnes (pondéré) si le revenu est diffusé (minimum général de 40). Pour les couples de même sexe, population totale minimale de 10 000 personnes. Pour les données de revenu, avoir aussi au moins 40 ménages privés (valeur pondérée).	On peut trouver la population d'une subdivision de recensement (SDR) ou d'un secteur de recensement (SR) dans les fichiers géographiques de suppression. Ces fichiers précisent également le nombre de ménages par SDR et SR.
7) Variables exprimées en dollars (revenu, par exemple) utilisées dans les modèles ou les résultats descriptifs	RÈGLE DE L'HOMOGENÉITÉ : L'intervalle doit représenter au moins X % du maximum; RÈGLE DE PRÉDOMINANCE : le maximum doit représenter moins de Y % de la somme.	Dans le cas des variables continues exprimées en dollars, on applique les deux critères en suivant la syntaxe disponible en SAS, STATA et SPSS



Règles de l'Enquête sur le milieu de travail et les employés

Les règles de l'EMTE diffèrent des règles principales

Exemples de différences:

- Fréquence minimale de 10 pour le fichier employeurs
- Tous les tableaux croisés doivent être pondérés
- Les résultats de modèles doivent être pondérés (sauf exception pour certaines méthodes et certains logiciels)



Règles de l'Enquête longitudinale auprès des immigrants du Canada

L'ELIC a certaines règles particulières:

- Fréquence minimale de 10
- Tableaux croisés pondérés seulement (sauf exception)

Règles des enquêtes post-censitaires

Les enquêtes post-censitaires comprennent:

- Enquête auprès des peuples autochtones
- Enquête sur la diversité ethnique
- Enquête sur la participation et les limitations d'activités

Elles ont des règles particulières:

- Fréquence minimale de 10
- Tableaux croisés pondérés seulement (sauf exception)
- Arrondissement des résultats descriptifs à la base 10



Réduction des risques de divulgation grâce au contrôle de la confidentialité

- Les chercheurs vérifient leurs résultats avant de les soumettre au contrôle de la confidentialité
- L'analyste du CDR passe en revue tous les résultats et les documents de soutien avant leur diffusion
- En temps normal, le contrôle de la confidentialité peut prendre jusqu'à trois jours; toutefois, si des problèmes se posent, le processus peut être plus long
- En général les résultats exploratoires ne sont pas soumis au contrôle de la confidentialité

Références sur la confidentialité

- Subcommittee on Disclosure Limitation Methodology, FCSM. (Second version, 2005). Report on Statistical Limitation Methodology. Statistical Policy Working Paper 22, Office of Management and Budget, Washington, D.C.
- Willenborg, L. and de Waal, T. (2000). Elements of Statistical Disclosure Control. Lecture Notes in Statistics, Springer Verlag.
- Willenborg, L. and de Waal, T. (1996). Statistical Disclosure Control in Practice. Lecture Notes in Statistics, Springer Verlag.
- Doyle, P., Lane, J., Theeuwes J. and Zayatz, L., eds. (2001). Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland.
- UN Economic Commission for Europe. (2006). *Managing Statistical Confidentiality and Microdata Access – Principles and Guidelines of Good Practice*. United Nations, Geneva.
- Hundepool, A. et coll. (2009). *Handbook on Statistical Disclosure Control, Version 1.1*. ESSNet SDC.



Pour en apprendre davantage

Site internet du Programme des CDR :

<http://www.statcan.gc.ca/rdc-cdr/index-fra.htm>

Coordonnées :

Jean-Louis Tambay : Jean-Louis.Tambay@statcan.gc.ca

Denis Gonthier : Denis.Gonthier@statcan.gc.ca