# Methodology of the Survey of Electronic Commerce and Technology 2005 (SECT)

## 1. Introduction

The Survey of Electronic Commerce and Technology 2005 (SECT) is an annual survey in its seventh year. It collects information on communication and technology such as the use of computers, Internet and web sites, as well as the use of Internet to do electronic commerce from a sample of Canadian enterprises.

The collection began in November 2005 and data for the reference year 2005 was published in April 2006. The data are collected for the 12 month fiscal period for which the final day occurs on or between January 1, 2005, and December 31, 2005.

## 2. Coverage

The sample used for this survey covers most industrial sectors. These are described using the North American Industrial Classification System (NAICS). Some sectors are excluded such as:

**A)** **Sector 11 Sub-sector 111, 112, 114, 1151 and 1152** (Crop and Animal Production Industries, Fishing, hunting and Trapping industries, Support Activities for Crop and Animal Production industries),
**B)** **Sector 23 Sub-sector 238** (Construction – Specialist contractors),
**C)** **Sector 91 Sub-sector 913** (Local Governments),
**D)** **Sector 55 Sub-sector 551114** (Head office),
**E)** **Sector 81 Sub-sector 814** (Private households).

## 3. Survey Frame and Target Universe

The frame consists primarily of the Business Register (**BR**) developed by Statistics Canada. The sampling unit is the enterprise. For more information on the Business Register and the sampling unit, refer to Cuthill (1998).

An administrative list is also used to cover the public sector. This list is provided and maintained for the needs of the survey by the Science, Innovation and Electronic Information Division (SIEID) at Statistics Canada. These units are sampled with certainty.

Because of the dynamic nature of businesses and/or units missed by the frame used, some units may be added once the sample has been selected to obtain a better coverage for the desired reference year. These units are sampled with certainty.

The initial sampling frame contains approximately

# Méthodologie de l'enquête sur le commerce électronique et la technologie 2005 (ECET)

## 1. Introduction

L'enquête sur le commerce électronique et la technologie 2005 (ECET) est une enquête annuelle qui en est à sa septième année d'existence. Elle collecte de l'information sur les communications et la technologie tels l'utilisation de l'ordinateur, l'Internet et les sites Web, ainsi que l'utilisation de l'Internet à des fins de commerce électronique auprès d'un échantillon d'entreprises canadiennes.

Les envois postaux ont débutés en novembre 2005 et des chiffres pour l'année 2005 ont pu être publiés dès avril 2006. On collecte l'information pour l'exercice financier de 12 mois se terminant entre le 1er janvier 2005 et le 31 décembre 2005.

## 2. Couverture

L'échantillon utilisé pour cette enquête couvre à peu près tous les secteurs industriels. Ceux-ci sont décrits en utilisant la convention connue sous le Système de classification industriel de l'Amérique du Nord (SCIAN). Quelques secteurs sont exclus tels :

**A)** **Secteur 11 sous-secteurs 111, 112, 114, 1151 et 1152 (**Industrie de la production animale et agricole, Industrie de la pêche, de la chasse et du piégeage, Activités de soutien à l'industrie de la production animale et agricole),
**B)** **Secteur 23 sous-secteur 238** (Construction - Entrepreneurs spécialisées),
**C)** **Secteur 91 sous-secteur 913** (Administrations locales),
**D)** **Secteur 55 sous-secteur 551114** (Bureaux-Chefs),
**E)** **Secteur 81 sous-secteur 814** (Ménages privés).

## 3. Base de sondage et population cible

La base de sondage est principalement formée du Registre des entreprises (**RE**) développé et maintenu à Statistique Canada. L'unité d'échantillonnage choisie est l'entreprise. Pour plus d'information sur le registre des entreprises et l'unité d'échantillonnage, se rapporter à Cuthill (1998).

Une liste administrative est également utilisée pour couvrir le secteur public. Cette liste est fournie et maintenue pour les besoins de l'enquête par la division des sciences, de l'innovation et de l'information électronique (DSIIE) de Statistique Canada. Ces unités sont échantillonnées avec certitude.

Étant donné la nature dynamique des entreprises et/ou des unités manquées sur la base de sondage utilisée, des unités peuvent être ajoutées une fois l'échantillon tiré afin d'obtenir une meilleure couverture pour l'année de référence voulue. Ces unités ajoutées sont échantillonnées avec certitude.

La base de sondage initiale compte environ 1 832 000 entreprises.

1,832,000 enterprises.

## Exclusions

Once the new universe is constructed, all units with income less than a certain limit are eliminated from the frame. We consider these units to have a negligible impact on electronic commerce and are of less interest. The exclusion allows us to reduce the response burden of small units.

The limit that delineates the out-of-scope units is determined as a function of industrial sector (NAICS2).

For the following NAICS2 sectors, we exclude from the target population the enterprises having a gross business income less than $100,000:

Sector 11 : Agriculture, Forestry, Fishing and Hunting
Sector 23 Construction
Sector 48-49 : Transportation and Warehousing
Sector 53 : Real Estate and Rental and Leasing
Sector 54: Professional, Scientific and Technical Services
Sector 56: Administrative and Support, Waste Management and Remediation Services
Sector 61 : Educational Services
Sector 62 : Health Care and Social Assistance
Sector 71 : Arts, Entertainment and Recreation
Sector 72 : Accommodation and Food Services
Sector 81 : Other Services (except Public Administration)

For the following NAICS2 sectors, we exclude from the target population the enterprises having a gross business income less than $250,000:

Sector 21 : Mining and Oil and Gas Extraction
Sector 22 : Utilities
Sector 31-33 : Manufacturing
Sector 41 : Wholesale Trade
Sector 44-45 : Retail Trade
Sector 51 : Information and Cultural Industries
Sector 52 : Finance and Insurance
Secror 55 : Management of Companies and Enterprises

It should be noted that these fixed boundaries allow a coverage of at least 95% of gross business revenue by industrial sector.

After exclusion, the sampling frame contains approximately 682,000 enterprises. This frame is our target population.

## 4. Sampling

The sampling consists of stratification, allocation and sample selection that are described in the following text.

## Stratification and Allocation

## Exclusions

Une fois la base de sondage établie, les unités du secteur privé ayant un revenu inférieur à une certaine limite sont éliminées de la base. On considère que ces unités ont un impact négligeable sur le commerce électronique et sont de moindre intérêt. L'exclusion permet aussi de réduire le fardeau de réponse des petites unités.

La limite inférieure déterminant les unités dans le champ de l'enquête est construite en fonction du secteur industriel (SCIAN2).

Pour les secteurs SCIAN2 suivants, on exclut de la population cible les entreprises du secteur privé ayant un revenu brut de moins de 100 000$ :

Secteur 11 : Agriculture, foresterie, pêche et chasse
Secteur 23 : Construction
Secteurs 48-49 Transport et entreposage
Secteur 53 : Services immobiliers et services de location et de location a bail
Secteur 54 : Services professionnels, scientifiques et techniques
Secteur 56 : Services administratifs, services de soutien, services de gestion des déchets et services d'assainissement
Secteur 61 : Services d'enseignement
Secteur 62 : Soins de santé et assistance sociale
Secteur 71 : Arts, spectacles et loisirs
Secteur 72 : Hébergement et services de restauration
Secteur 81 : Autres services, sauf les administrations publiques

Pour les secteurs SCIAN2 suivants, on exclut de la population cible les entreprises du secteur privé ayant un revenu brut de moins de 250,000$ :

Secteur 21 : Extraction minière et extraction de pétrole et de gaz
Secteur 22 : Services Publics
Secteurs 31-33 : Fabrication
Secteur 41 : Commerce de gros
Secteurs 44-45 : Commerce de détail
Secteur 51 : Industrie de l'information et industrie culturelle
Secteur 52 : Finance et assurances
Secteur 55 : Gestion de sociétés et d'entreprises

Il est à noter que ces bornes fixes permettent une converture du revenu par secteur industriel d'au moins 95%.

Après exclusion, la base de sondage échantillonnale compte environ 682 000 entreprises. Cette base de sondage correspond à notre population cible.

## 4. Échantillonnage

L'échantillonnage comprend la stratification, la répartition et la sélection de l'échantillon qui sont décrites dans le texte qui suit.

## Stratification et répartition

Tout d'abord, quelques unités pour lesquelles on s'attend à de très grandes ventes par Internet ont été identifiées. Ces unités prédéterminées ont été sélectionnées avec certitude et ont pu

First, some units for which we expect very large sales over the Internet were identified. These predetermined units were to be selected with certainty and thus were removed from the stratification and allocation process described below.

The remaining units on the frame were first stratified by NAICS at the level required for estimation. Then, within each industrial level, we built four strata by size defined using the number of employees:

Three take-some strata for which the sampling is conducted using a probability of selection:

Stratum 1 : 0 to 19 employees
Stratum 2 : 20 to 99 employees
Stratum 3 : 100 to 499 employees

One take-all stratum for which the units are sampled with certainty

Stratum 4 : 500 employees and more

These groupings follow the definition of small, medium and large enterprises used at estimation

The sample allocation was done using the Neyman method (Cochran, 1977). We took into account two types of variability by stratum: the variance of the gross business income and the estimated variance of sales over Internet from the previous year. The available sample size was 19,000 units that were first allocated in differents NAICS groupings to better represent sectors more favourable for electronic commerce.

Once the stratification and the allocation were done, we increased the sample size in some strata when necessary in order to obtain a minimum sampling fraction of 1% and a minimum of five units by stratum when possible. The next step is to select the sample of enterprises.

## Selection

All predetermined units and all units in the take-all strata were selected with certainty, while a random sample was selected in the take-some strata under the constraint of maximizing the overlap with the previous year's sample. The Kish and Scott method (1971) was used and a global overlap of almost 71% with the last sample was obtained.

## 5. Collection and Data Editing

A questionnaire was mailed to enterprises and respondents were encouraged to complete and return it.

At data collection, some edits were applied to each questionnaire such as rules of consistency and historical edits. For more details on the edit rules, see Uhrbach (2005).

être exclues du processus de stratification et de répartition décrit ci-dessous.

Les unités restantes de la base ont tout d'abord été stratifiées selon le SCIAN suivant le niveau désiré pour les estimations. Ensuite, à l'intérieur de chaque niveau industriel, on a créé quatre strates de taille définie selon le nombre d'employés:

Trois strates à tirage partiel pour lesquelles l'échantillonnage se fait selon une probabilité de sélection :

Strate 1 : 0 à 19 employés
Strate 2 : 20 à 99 employés
Strate 3 : 100 à 499 employés

Une strate à tirage complet pour laquelle l'échantillonnage se fait avec certitude:

Strate 4 : 500 employés et plus.

Ces groupements suivent les définitions de petites, moyennes et grandes entreprises utilisées à l'estimation.

La répartition de l'échantillon a été faite selon la méthode de Neyman (Cochran, 1977). Nous avons tenu compte de deux types de variabilité par strate : la variance du revenu brut de l'entreprise et la variance estimée des ventes par Internet de l'année précédente. La taille d'échantillon permise était de 19 000 unités que nous avons redistribuées au départ dans différents groupements de SCIAN afin de mieux représenter les secteurs plus propices au commerce électronique.

Une fois la stratification et la répartition effectuées, nous avons augmenté la taille de l'échantillon dans certaines strates si nécessaire de sorte à obtenir une fraction d'échantillonnage minimale de 1% et un minimum de cinq unités. La prochaine étape consiste à sélectionner l'échantillon d'entreprises.

## Sélection

Toutes les unités prédéterminées et toutes les unités dans les strates à tirage complet ont été échantillonnées avec certitude alors qu'un échantillon aléatoire a été tiré dans les strates à tirage partiel sous la contrainte de maximiser le chevauchement avec l'échantillon de l'année précédente. La méthode de Kish et Scott (1971) a alors été utilisée et un chevauchement global de près de 71% a été obtenu avec l'échantillon précédent.

## 5. Collecte et traitement des données

Un questionnaire a été envoyé par la poste aux entreprises invitant le répondant à le retourner dûment rempli.

À la saisie des données, des règles de vérification ont été appliquées à chaque questionnaire, telles des règles de cohérence et de vérification historique. Pour plus de détails sur les règles de vérification, consulter Uhrbach (2005).

Les unités n'ayant pas répondu ont fait l'objet de suivis postaux et

Units that had not responded or had answered incorrectly were subject to mail and fax follow-up to ensure the data was obtained. Also, some follow-ups were done by phone in order to increase the response rate and improve the representativity of the sample.

Follow-ups were done on questionnaires received to get data not reported, to correct inconsistencies in the data or to validate/correct data significantly different with historical data.

Finally, we prioritized the follow-ups by taking into account the response rate by industrial sector, the size of the enterprise, the importance of the missing variables and the kind of inconsistencies on the questionnaire.

The definition of response rate varies depending on the needs. We will give here the response rate based on responding units among units where a questionnaire was sent.

Units sampled: 19,434 enterprises
Units sent out for data collection: 18,031 enterprises
Responding units: 12,583 enterprises
Response rate : 70%

Some units selected are not sent for data collection. These are units where their status changed since the frame was created and/or are errors on the frame such as duplicates, out-of-business or out-of-scope. There is no interest to send these units for collection.

## 6. Outlier Detection

Outlier detection was done on the variable “Sales over the Internet” as collected in the 2005 survey. We also did outlier detection on the year over year difference between sales over the Internet in 2005 and in 2004. In both cases, the detection was made within groups formed according to the private/public sector and the industrial sector (NAICS-2 level) if there were at least 10 units in the group. Otherwise, the detection was done by private/public sector only. A method using the distance between observations was used (Nobrega, 1998).

For outlier detection on sales over the Internet for 2005, more than 50 units were detected as outliers. These units were analyzed and corrected as necessary. The units that are outliers and correct were promoted to a take-all stratum in order to represent only themselves. We consider that these units are misclassified during the sampling and do not correctly represent other units in the stratum. The selection probability for residual units was then recomputed.

For outlier detection on the year over year difference between sales over the Internet in 2004 and in 2005, about 20 units were detected as outliers. Most of these units had already been detected as outliers because of the value of their sales over the Internet. Those units

par télécopieur afin d’obtenir leurs réponses. Certains suivis par téléphone ont également été faits afin d’améliorer le taux de réponse et la représentativité de l’échantillon.

Des suivis ont été faits sur les questionnaires reçus afin d’obtenir des données non-rapportées, de corriger des données incohérentes ou encore de valider/corriger des données significativement différentes lorsque comparées aux données historiques.

Enfin, nous avons priorisé les suivis en tenant compte du taux de réponse par secteur industriel, de la taille de l’entreprise, de l’importance des variables manquantes et du type d’incohérences sur le questionnaire.

La définition d’un taux de réponse diffère selon les besoins. On donnera ici un taux de réponse basé sur le nombre d’unités répondantes parmi les unités envoyées à la collecte.

Unités échantillonnées : 19 434 entreprises
Unités envoyées à la collecte : 18 031 entreprises
Unités répondantes : 12 583 entreprises
Taux de réponse : 70%

Certaines unités échantillonnées ne sont pas envoyées à la collecte. Il s’agit d’unités dont le statut a changé depuis la création de la base de sondage et/ou d’erreurs sur la base de sondage telles des unités en double, plus en affaire ou hors du champ de l’enquête. Il n’est d’aucun intérêt d’envoyer ces unités à la collecte.

## 6. Détection de données aberrantes

Une détection des données aberrantes a été faite sur la variable des ventes par Internet telle que rapportée dans l’enquête de 2005. Il y a aussi eu une détection de valeurs aberrantes pour la différence entre les ventes par Internet rapportées en 2005 et rapportées en 2004. Dans les deux cas, la détection a été faite à l’intérieur de groupes formés selon le secteur privé/public et le secteur industriel (SCIAN de niveau 2) s’il y avait au moins 10 unités dans le groupe. Sinon, on a fait la détection par secteur privé/public seulement. Une méthode basée sur les écarts entre les observations a été utilisée (Nobrega, 1998).

Pour la détection de valeurs aberrantes pour la variable des ventes par Internet pour 2005, plus de 50 unités ont été détectées aberrantes. Ces données ont ensuite été vérifiées et corrigées au besoin. Les unités trouvées aberrantes et valides ont été promues dans une strate à tirage complet afin de ne représenter qu’elles-mêmes. On considère ces unités mal classifiées lors de l’échantillonnage et ne représentant pas correctement les autres unités de la strate. La probabilité de sélection des unités résiduelles a été recalculée.

Pour la détection de valeurs aberrantes pour la différence entre les ventes par Internet rapportées en 2004 et en 2005, environ 20 unités ont été détectées aberrantes. La majorité de ces unités avaient déjà été détectées aberrantes à cause de la valeur de leurs ventes par Internet. Ces unités avaient donc déjà été traitées (voir paragraphe ci-dessus). Pour les unités restantes, nous n’avons apporté aucun traitement : ces unités avaient un

had already been treated (see above paragraph). For remaining units, no treatment has been done: those units had a weight close to 1 and represented almost only themselves.

## 7. Edit and Imputation

Once the survey collection was closed, some records remained incomplete and/or inconsistent. The missing and/or inconsistent fields on these records were imputed. Globally, 6% of the fields were imputed due to missing data while 0.1% of the fields were imputed due to inconsistencies. Only partial questionnaires were imputed. In the case of total non-response, no imputation was performed. We simply reweighted responding units at estimation (see section 8. Estimation).

Many imputation methods were used: deterministic imputation, imputation using administrative data, historical imputation and donor imputation.

**Deterministic imputation** was used when answers from questions related to the question needing imputation lead to only one possible answer. 0.7% of the fields were imputed in this matter.

**Imputation using administrative data** was used to impute the question referring to the number of employees by using the number of employees available on the BR. 4.6% of the fields referring to the number of employees were imputed.

**Historical imputation** was used to impute some categorical variables that are stable over time when the enterprise positively responded the year before. Also, for total sales over the Internet, we imputed historically in cases where the enterprise reported doing sales over the Internet without reporting the amount. For those cases, we imputed using last year's value adjusted for the year over year trend. The trend was calculated within each industrial sector if at least 10 enterprises did sales over internet for the 2 years, otherwise the trend was calculated within each private/public sector. Only 64 fields were imputed using historical information.

**Donor imputation** was finally used in the remaining cases to replace missing or incoherent values with those of the nearest respondent according to characteristics such as size, industrial classification and key variables from the questionnaire. We also checked to be sure that the imputed values did not affect the questionnaire's consistency. Imputation was conducted within homogeneous groups, the initial imputation group corresponding to the stratum. If there were not at least 10 potential donors and 25% of donors in a group, or if imputation from all available donors would result in questionnaire inconsistencies, we moved to a more aggregated imputation group in the following order:
NAICS-3 level and size grouping;
NAICS-3 level;
NAICS-2 level and size grouping;
NAICS-2 level.

poids très près de 1 et ne représentaient en fait presque qu'elles-mêmes.

## 7. Vérification et Imputation

Une fois l'enquête terminée, il restait certains enregistrements toujours incomplets et/ou incohérents. Les champs manquants et/ou incohérents de ces enregistrements ont été imputés. Globalement, 6% des champs ont dû être imputés parce que le champ était manquant et environ 0.1% des champs parce qu'il y avait incohérence entre les champs. Seuls les questionnaires partiels ont été imputés. Dans le cas d'une non-réponse totale, aucune imputation n'a été faite. On a plutôt repondéré à l'estimation les unités répondantes (voir section 8. Estimation).

Plusieurs types d'imputation ont été utilisés, soit l'imputation déterministe, l'imputation par source administrative, l'imputation historique et l'imputation par donneur.

**L'imputation déterministe** a été effectuée lorsque les réponses aux questions reliées à la question à imputer ne laissaient qu'un seul choix de réponse. 0,7% des champs ont ainsi été imputés.

**L'imputation par source administrative** a été effectuée pour la question portant sur le nombre d'employés en utilisant le nombre d'employés disponible sur le registre des entreprises. 4,6% des champs portant sur le nombre d'employés ont été imputés.

**L'imputation historique** a été utilisée pour imputer certaines variables catégoriques stables dans le temps lorsque l'entreprise avait répondu dans l'affirmative l'année précédente. De plus, les ventes totales par Internet ont été imputés historiquement lorsque l'entreprise a mentionné faire des ventes par Internet sans en donner la valeur. Dans ce cas, on a imputé par la valeur de l'année précédente si disponible, en ajustant par la tendance. La tendance a été calculée à l'intérieur de chaque secteur industriel si au moins 10 entreprises dans le secteur avaient fait des ventes par Internet pour les 2 années, sinon la tendance a été calculée à l'intérieur de chaque secteur privé/public. Seulement 64 champs ont été imputés en utilisant l'information historique.

**L'imputation par donneur** a finalement été effectuée dans les autres cas en remplaçant les valeurs manquantes ou incohérentes par celles du plus proche répondant selon certaines caractéristiques telles la taille, la classification industrielle et les variables-clé du questionnaire. De plus, on s'est assuré que le donneur permettait de respecter la cohérence entre les champs imputés et les champs rapportés du receveur. L'imputation a été exécutée à l'intérieur de groupes homogènes, le groupement initial correspondant à la strate. Si on n'avait pas au moins 10 donneurs potentiels et 25% de donneurs par groupe ou encore, si aucun donneur disponible ne permettait l'imputation tout en respectant les règles de validation du questionnaire receveur, on passait à un groupe d'imputation plus agrégé et dans l'ordre suivant:
SCIAN de niveau 3 et les groupes de taille;
SCIAN de niveau 3;
SCIAN de niveau 2 et les groupes de taille;
SCIAN de niveau 2.
Secteur privé/public.

Private/Public Sector.

Note that outlier enterprises were excluded from the donor pool. When imputation was done, we adjusted the sales value over the Internet by the ratio of imputed and donor's revenue. 5.6% of the fields were imputed by donors.

When we could not find a donor for an enterprise, it was manually imputed. This situation did not happen this year. Finally, when imputation was completed, we reapplied the initial edit rules to assure the consistency of all the questionnaires going into the estimation process. Imputation flags were created to keep information about imputed fields. Also, outlier detection was performed again on sales over the Internet as collected in 2005 and on the year over year difference between sales over the Internet in 2005 and in 2004 in order to detect outliers that could have been created during the imputation.

## 8. Estimation

Statistics Canada's Generalized Estimation System (GES) was used (see 2001 GES). The estimation was done in two phases: the first phase sample was the initial sample and the second phase sample was the respondents. The same stratification was used at both the first and the second phases.

Three types of estimates were produced:

1) In the case of **percentage variables (*P*)**, a ratio was used to derive an estimate.

$$\hat{P}_d = \frac{\sum_s w_i z_i p_i(d)}{\sum_s w_i z_i} \text{ where } p_i(d) = \begin{cases} p_i \; if \; i \varepsilon d \\ 0 \; otherwise \end{cases}$$

2) In the case of **categorical variables (*C*)**, again a ratio was used.

$$\hat{C}_d = \frac{\sum_s w_i z_i c_i(d)}{\sum_s w_i z_i} \text{ where}$$

$$c_i(d) = \begin{cases} 1 if \; i \varepsilon d \; and \; the \; category \; was \; chosen \\ 0 \; otherwise \end{cases}$$

3) In the case of **numerical variables (Y)**, the usual estimator of the total was used.

$$\hat{Y}_d = \sum_s w_i y_i(d) \text{ where } y_i(d) = \begin{cases} y_i \; if \; i \varepsilon d \\ 0 \; otherwise \end{cases}$$

Notons que les questionnaires avec données aberrantes étaient exclus du bassin de donneurs. Une fois l'imputation effectuée, on a ajusté la variable des ventes par Internet par le ratio des revenus du receveur et du donneur. 5,6% des champs ont été imputés par donneur.

Dans les cas où on ne peut trouver un donneur pour une entreprise, ces unités sont imputées manuellement. Cette situation n'est pas survenue cette année. Enfin, une fois l'imputation terminée, les règles de vérification initiales ont été réappliquées afin de s'assurer de la cohérence de tous les questionnaires utilisés à l'estimation. Des drapeaux d'imputation ont été créés afin de garder l'information des variables imputées. De plus, la détection des données aberrantes a été refaite sur les ventes par Internet rapportées en 2005 et sur la différence entre les ventes par Internet rapportées en 2005 et celles rapportées en 2004 de sorte à détecter les valeurs aberrantes qui auraient pu être créées lors de l'imputation.

## 8. Estimation

Le système généralisé d'estimation (SGE) de Statistique Canada a été utilisé (voir 2001 SGE). L'estimation s'est fait en deux phases : l'échantillon de première phase étant l'échantillon initial et l'échantillon de deuxième phase, les répondants. La même stratification a été utilisée en première et deuxième phases.

Trois types d'estimations ont été produits :

1) Dans le cas des **variables de pourcentage (*P*)**, un quotient a été utilisé pour produire les estimations.

$$\hat{P}_d = \frac{\sum_s w_i z_i p_i(d)}{\sum_s w_i z_i} \text{ où } p_i(d) = \begin{cases} p_i \; si \; i \varepsilon d \\ 0 \; si \; non \end{cases}$$

2) Dans le cas des **variables catégoriques (*C*)**, à nouveau un quotient a été utilisé.

$$\hat{C}_d = \frac{\sum_s w_i z_i c_i(d)}{\sum_s w_i z_i} \text{ où}$$

$$c_i(d) = \begin{cases} 1 si \; i \varepsilon d \; et \; la \; catégorie \; a \; été \; choisie \\ 0 \; si \; non \end{cases}$$

3) Dans le cas des **variables numériques (Y)**, l'estimateur habituel du total a été utilisé.

$$\hat{Y}_d = \sum_s w_i y_i(d) \text{ où } y_i(d) = \begin{cases} y_i \; si \; i \varepsilon d \\ 0 \; si \; non \end{cases}$$

La variable $w_i$ représente le poids final de l'unité i après

The variable $w_i$ represents the final weights of the unit i after reweighting to take into account the non-response. The variable $z_i$ is the auxiliary variable that may be revenue, the number of employees or others depending on the variable being estimated. This variable, if used, allows us to produce economically weighted estimates which give more weight to large units.

For formulas for variance estimation of a two-phase design for each type of variable (*P*, *C* and *Y*), please refer to Arcaro (1998).

**Calculation of CV**

The coefficient of variation (CV) is computed using the ratio:

$$CV(\hat{Y}(d)) = \frac{\sqrt{\hat{V}(\hat{Y}(d))}}{\hat{Y}(d)}$$

where the numerator represents the estimate's standard deviation. Variable Y may represent any of the types of variables already discussed. However, in cases of percentage or categorical variables, we modified the CV calculation by using Y(d)=0.5. This way, we avoid getting very small or very large CVs due to Y(d) being close to 1 or close to 0.

This coefficient tries to give a relative measure of the error made when using a sample instead of using a census to derive an estimate about the whole population.

## 9. Confidentiality

Some confidentiality rules were used to suppress any information that might lead to disclosure of the data supplied by a respondent. These rules allow Statistics Canada to comply with its mandate of non-disclosure of information supplied by respondents. The rules themselves are confidential and are not available for consultation.

## 10. Sampling Error and Non-Sampling Error

The difference between an estimate based on sample data and the value obtained by surveying the entire population is called the sampling error. This difference varies with sample size, variability of the variable of interest, sampling design, and estimation method. In general, the larger a sample, the smaller its sampling error. If the population is very heterogeneous, a larger sample size is required to produce a reliable estimate.

The sampling error is measured by a quantity known as the standard deviation. The latter indicates the

repondération pour tenir compte de la non-réponse. La variable $z_i$ est une variable auxiliaire qui peut être le revenu, le nombre d'employés ou autre selon la variable estimée. Des estimations sont produites avec et sans cette variable auxiliaire. Cette variable permet de dériver des estimations qu'on appelle économiquement pondérées en donnant plus de poids aux unités de grandes tailles.

Pour ce qui est des formules d'estimation de variance d'un plan à deux phases pour chacune des catégories de variable (*P*,*C* et *Y*), il faut se référer à Arcaro (1998).

**Calcul du CV**

Le coefficient de variation (CV) est calculé à l'aide du quotient:

$$CV(\hat{Y}(d)) = \frac{\sqrt{\hat{V}(\hat{Y}(d))}}{\hat{Y}(d)}$$

où le numérateur représente l'écart-type échantillonnal de l'estimation. La variable Y peut représenter chacun des types de variables discutés préalablement. Par contre, dans le cas de pourcentages et de variables catégoriques, on a modifié le calcul du CV en utilisant Y(d)=0.5. On évite ainsi d'obtenir de très petits ou grands CV reliés au fait que Y(d) soit très près de 1 ou très près de 0.

Ce coefficient tente de donner une mesure relative de l'erreur commise lorsqu'on a recours à un échantillon au lieu de produire une statistique à l'aide de l'ensemble de la population.

## 9. Confidentialité

Certaines règles de confidentialité ont été utilisées pour supprimer toute information qui pourrait mener à la divulgation des données fournies par un répondant. Ces règles permettent à Statistique Canada de respecter son mandat de non-divulgation d'information fournie par les répondants. Les règles elles-mêmes sont confidentielles et ne sont pas disponibles pour consultation.

## 10. Erreur d'échantillonnage et non-due à l'échantillonnage

La différence entre l'estimation produite à partir de données échantillonnées et de données recensées est appelée erreur d'échantillonnage. Cette différence varie plus ou moins selon la taille de l'échantillon, la variabilité de la variable d'intérêt, le plan de sondage et la méthode d'estimation. En général, un échantillon plus grand produit une erreur d'échantillonnage plus petite. Si la population est très hétérogène, une taille d'échantillon plus grande est requise pour produire une estimation fiable.

L'erreur d'échantillonnage est mesurée par une quantité appelée écart-type. Cette quantité mesure la variabilité anticipée de l'estimation produite si on fait un échantillonnage répété. La vraie valeur de l'écart-type est inconnue mais peut être estimée à partir

expected variability of the estimate that would be produced if we sampled repeatedly. The actual value of the standard deviation is unknown, but it can be estimated from the sample.

Another measure of precision is the coefficient of variation (CV). The CV is simply the standard deviation expressed as a percentage of the estimate. Hence it is a relative measure of precision and can be used for comparisons across industries or provinces. The smaller the CV, the more reliable the estimate.

As well as sampling error, there are non-sampling errors such as frame problems, response errors, data capture errors, etc. Although every effort is made to keep such errors to a minimum, they always exist. They are not taken into account in computing the CV. Measures such as response rate, coverage rate, imputation rate and non-response studies (Duval, 2005) can be used as indicators of the possible extent of non-sampling errors.

Here are some results of the response rate among the 19,434 enterprises sampled:

Questionnaires completed: 37%
Questionnaires partially completed: 24%
No response before deadline: 25%
Unable to locate: 10%
Out-of-scope or out-of-business: 4%
Refusal: 0%

When the estimates are published, a scale distinguishes between the various qualities of accuracy. It combines the effect of sampling (using the CV) and the imputation rate (each imputed value adds to the uncertainty of the results). The scale is presented in Table 6.

**Table 6**
## Quality indicator interpretation

| | Imputation rate | | | |
|---|---|---|---|---|
| CV | 0.00 - 0.10 | 0.10 - 0.33 | 0.33 - 0.60 | 0.60 - +++ |
| 0.00 - 0.05 | A | B | C | F |
| 0.05 - 0.10 | B | C | D | F |
| 0.10 - 0.15 | C | D | E | F |
| 0.15 - 0.25 | D | E | F | F |
| 0.25 - 0.50 | E | F | F | F |
| 0.50 - +++ | F | F | F | F |

A: Excellent B: Very good C: Good
D: Acceptable E: Use with caution F: Unpublishable

## 11. References

(2001). Generalized Estimation System. Internal Statistics Canada document, October 2001.

Arcaro C. (1998). GES Estimation Specifications for



de l'échantillon.

Une deuxième mesure de précision est le coefficient de variation (CV). Ce coefficient est simplement l'écart-type exprimé en pourcentage de la valeur de l'estimation. Il donne donc une mesure de précision relative et comparable entre différentes industries ou provinces. Notons qu'un plus petit CV indique une plus grande fiabilité de l'estimation.

En plus de l'erreur d'échantillonnage, il existe des erreurs non-dues à l'échantillonnage telles des problèmes de base de sondage, des erreurs de réponses, des erreurs lors de l'encodage des réponses, etc., sur lesquelles on tente de conserver un contrôle des plus stricts. Néanmoins, celles-ci existent toujours et ne sont pas comptabilisées lorsque l'on calcule le coefficient de variation. Certaines mesures telles que des taux de réponse, de couverture, d'imputation et des études sur la non-réponse (Duval, 2005) peuvent être utilisées comme indicateurs du niveau potentiel des erreurs non-liées à l'échantillonnage.

Voici des résultats concernant le taux de réponse des 19 434 entreprises échantillonnées:

Questionnaires complétés : 37%
Questionnaires partiellement complétés : 24%
Pas de réponse avant la date limite de l'enquête : 25%
Pas de contact possible : 10%
Hors du cadre de l'enquête ou plus en affaire : 4%
Refus : 0%

Lors de la publication des estimations, une échelle permet de distinguer entre les différentes qualités de précision. Celle-ci combine l'effet dû à l'échantillonnage (à l'aide du CV) et le taux d'imputation (chaque imputation ajoute à l'incertitude des résultats). L'échelle utilisée est reproduite au tableau 6.

**Tableau 6**
## Interprétation de la côte de qualité

| | Taux d'imputation | | | |
|---|---|---|---|---|
| CV | 0.00 - 0.10 | 0.10 - 0.33 | 0.33 - 0.60 | 0.60 - +++ |
| 0.00 - 0.05 | A | B | C | F |
| 0.05 - 0.10 | B | C | D | F |
| 0.10 - 0.15 | C | D | E | F |
| 0.15 - 0.25 | D | E | F | F |
| 0.25 - 0.50 | E | F | F | F |
| 0.50 - +++ | F | F | F | F |

A: Excellent B: Très bon C: Bon
D: Acceptable E: Utiliser avec réserve F: Non-publiables

## 11. Références

(2001). Système Généralisé d'Estimation. Document interne de Statistique Canada, Octobre 2001.

Arcaro C. (1998). GES Estimation Specifications for Two-Phase Sampling with Auxiliary Information, Document interne de

Two-Phase Sampling with Auxiliary Information, Internal Statistics Canada document, 1998.

Cochran William G., (1977). “Stratified Random Sampling”, *Sampling Techniques*, Wiley, pp. 99-101.

Cuthill I. (1998). The Statistics Canada Business Register. Internal Statistics Canada document, 1998.

Duval M-C. (2005). Étude de non-réponse pour l’enquête sur le commerce électronique 2004. , Internal Statistics Canada document, May 2005.

Kish L. and Scott A. (1971). Retaining Units after Changing Strata and Probabilities. Journal of the American Statistical Association, September 1971, 461-470

Nobrega K. (1998). Outlier Detection in Asymmetric Samples: A Comparison of an Inter-quartile Range Method and a Variation of a Sigma Gap Method. Statistical Society of Canada, 1998 Proceedings of the Survey Methods Section, June 1998.

Uhrbach M. (2005). New Edits for 2005. Internal Statistics Canada document, October 2005.

Statistique Canada, 1998.

Cochran William G., (1977). “Stratified Random Sampling”, *Sampling Techniques*, Wiley, pp. 99-101.

Cuthill I. (1998). Le registre des entreprises de Statistique Canada. Document interne de Statistique Canada, 1998.

Duval M-C. (2005). Étude sur les non-répondants de l’enquête sur le commerce électronique de 2004. Document interne de Statistique Canada, mai 2005.

Kish L. et Scott A. (1971). Retaining Units after Changing Strata and Probabilities. *Journal of the American Statistical Association*, September 1971, 461-470

Nobrega K. (1998). Outlier Detection in Asymmetric Samples: A Comparison of an Inter-quartile Range Method and a Variation of a Sigma Gap Method. *Statistical Society of Canada*, 1998 Proceedings of the Survey Methods Section, June 1998.

Uhrbach M. (2005). New Edits for 2005. Document interne de Statistique Canada, octobre 2005.